NOTE

This Microfiche meets the quality requirements of the American National Standard for Microfiche of Documents (ANSI PH5.9-1975). However, because of imperfections in the document from which it was made, it contains images that are illegible or difficult to read. The source document for this microfiche is the best copy that could be obtained prior to filming.

CIS Congressional Information Service, Inc.
Washington, D.C.

Congress of the United States
Office of Technology Assessment

MAPPING OUR GENES
Federal Genome Projects: How Vast? How Fast?

Contractor Reports, Volume 2

February 1988

The Mapping and Sequencing of Genomes: A Comparative Analysis of Methods, Benefits and Disbenefits, Stephen M. Mount, Department of Biological Sciences, Columbia University, New York, New York (October 1987)

Mapping the Human Genome: Experimental Approaches for Cloning and Ordering DNA Fragments, Richard M. Myers, Department of Physiology, University of California-San Francisco (October 1987)

Mapping and Sequencing the Human Genome in Europe, Peter A. Newmark, *Nature* Magazine, London, England (October 1987)

Application of Human Genome Mapping for the Global Control of Genetic Disease, Sir David J. Weatherall, Nuffield Department of Clinical Medicine, John Radcliffe Hospital, University of Oxford, Oxford, England (August 1987)

In Search of the "Ultimate Map" of the Human Genome: The Japanese Efforts, Akihiro Yoshikawa, Berkeley Roundtable on the International Economy, University of California-Berkeley, Berkeley, California (August 1987)

These are contractor reports used in preparing OTA's final assessment report. OTA makes them available for the use of readers desiring a more detailed or technical discussion of an issue than can normally be accommodated in our final report. As OTA contractor reports, they have not been reviewed or approved by the Technology Assessment Board. The findings and conclusions expressed are those of the authors and do not necessarily reflect the views of OTA, the assessment advisory panel or the Technology Assessment Board.

J952-8

This document can be ordered in either paper or microfiche format from:

National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road
Springfield, Virginia 22161
Telephone: (703) 487-4650
Telex: 89-9405

THE MAPPING AND SEQUENCING OF GENOMES:

A Comparative Analaysis of Methods, Benefits and Disbenefits.

Stephen M. Mount, Ph.D.
Room 922A Fairchild
Department of Biological Sciences
Columbia University
New York, NY 10027

CONTENTS

THE MAPPING AND SEQUENCING OF GENOMES: A COMPARATIVE ANALYSIS OF METHODS, BENEFITS AND DISBENEFITS.

I. Methods for obtaining physical maps of genomes, and the current status of such efforts. p. 4

Introduction. P. 4

Current status of maps and sequence information in seven well-studied species. p. 7
1. *Escherichia coli*, an enteric bacterium p. 7
2. *Saccharomyces cerevisiae*, baker's yeast p. 8
3. *Arabadopsis*, duckweed p. 9
4. *Caenorhabditis elegans*, a nematode worm P. 9
5. *Drosophila melanogaster*, a fruit fly p. 10
6. *Mus musculus*, the mouse p. 13
7. *Homo sapiens*, humans p. 13

II. Scientific arguments for a concerted effort to map the genomes of other species in parallel with the human genome. p. 14

Feasibility and special problems in various species. p. 14

1. Proportion of sequence likely to be coding p. 14
2. Unclonable sequences p. 15
3. Heterozygosity p. 15
4. Correlations with genetic and physical maps p. 15
5. Effects of size p. 16

Relative values of human and nonhuman maps and sequences. p. 16

Associated technologies which will greatly increase the value of the map or sequence. p. 20

III. Recommendations based on the above arguments. p. 22

Figure legends. p. 28

References to the literature. p. 29

Telephone consultants p. 31

I. Methods for obtaining physical maps of eukaryotic genomes, and the current status of such efforts.

INTRODUCTION

DNA, the genetic material, is organized in cells as an unbranched string of nucleotides. Each cell typically (there are many variations on this general condition, depending on the species) contains one or two copies of the entire set of genetic information present in that organism, which is known as the genome. In eukaryotes (organisms, including all multicellular animals and plants, in which the cells contain nuclei), the bulk of the DNA is present in the nucleus, and is organized into linear segments known as chromosomes. Relatively small amounts of DNA can also be found in organelles such as mitochondria and chloroplasts. Mitochondrial and chloroplast genomes, like bacterial genomes, often differ from the nuclear genome in being circular rather than linear. However, all DNA is unbranched. Thus, all genomes can be organized into one-dimensional maps. There are a variety of types of map (physical, genetic and cytological maps will be distinguished below), but each ultimately refers to the same one-dimensional sequence of nucleotides, and the order of points must be the same for all maps of a single genome. Maps can differ in the relative distances between two points, but they cannot differ in the order of corresponding points.

A physical map describes the location of physical features along the DNA. These features are typically the sites of cleavage by sequence-specific endonucleases (restriction enzymes) or sites where a particlular sequence can be detected using techniques involving hybridization. A physical map can be constructed in the complete absense of genetic or cytological data. However, such data is generally quite useful in the construction of a physical map, and knowledge of the location of genes is the purpose of constructing a physical map. The single nucleotide is the smallest unit of genetic information, so it is not possible to have more detailed information

than the nucleotide sequence. Thus, the nucleotide sequence of a region of the genome is the physical map of that region with the highest possible resolution, and a physical map with lower resolution (such as a map of restriction enzyme cleavage sites) can be deduced from the nucleotide sequence. It should be noted, however, that deduction of the location of genes from a nucleotide sequence (or any other physical map) is not entirely straightforward (see below).

A genetic map is the order of genes along a chromosome as determined by linkage. When genes are passed on to progeny, those genes which lie on different chromosomes will segregate independently, and are described as being completely unlinked. Those genes which lie on the same chromosome will have a greater chance of being transmitted together. This would occur invariably if it were not for recombination between homologous chromosomes during meiosis, the specialized cell division which occurs in the generation of gametes (sperm and egg cells). Genes which are closer together along the DNA within a single chromosome are more likely to be inherited together because it is less likely that they will be separated by a recombination event occurring between them. Such genes are referred to as being closely linked. The exact probabiltiy of two genes being separated by recombination gives a quantitative measure of how closely linked the two genes are, with a probability of 1% corresponding to one centiMorgan (cM). Recombination between two adjacent genes will generally also produce new combinations of more distant genes. In these ways, examination of the transmission of genes allows them to be ordered (see Figure 1).

Thus, a genetic map is generated by studying the transmission of genes. However, in order to follow the transmission of genes there must be at least two detectably different states (or alleles) of each gene. These are typically the normal form (or wild type) of the gene and a mutant form of the gene. Sometimes, however,

multiple forms of a gene which have no obvious effects on appearance (no phenotype) can be distiguished by other means. These distinct wild type forms are known as polymorphisms. A form of polymorphism which has proven especially useful in human genetic analysis is the restriction fragment length polymorphism or RFLP, in which variation between individuals in the length of particular restriction enzyme fragments is used as a genetic marker. RFLPs are analyzed in the same manner as other genetic markers (by linkage analysis), and are used in the construction of a genetic map, but have the characteristic of being directly tied to the physical map. As physical markers which can be followed genetically, they are therefore extremely useful in correlating genetic and physical maps.

A cytological map is one based on microscopic visualization of chromosomes. In some species, most notably Drosophila, large easily visualized chromosomes, often consisting of more than a thousand parallel copies of DNA, occur in some cells. These are known as polytene chromosomes. The visualization of polytene chromosomes (see Figure 2) allows the construction of cytological maps with a resolution in the range of 10-100 kb (kilobases; one kilobase is one thousand -- 10^3 -- nucleotides; 1 Mb or Megabase is one million -- 10^6 -- nucleotides). These maps have a resolution in the range of one to several genes. Polytene chromosomes do not exist in mammalian species, and the resolution of cytological maps (which are based on the examination of chromosomes during mitosis) is considerably less; estimates are in the range of 1 to 10 Mb (Funanage and Myoda, 1986; Goodenough, 1984).

In all cases, the cytological map is correlated with the physical and genetic maps by two principle methods. One is the hybridization of cloned DNA to chromosomes - the site of hybridization of labelled cloned DNA indicates the location of that DNA on the chromosome. The other method is analysis of breakpoints of deletions, inversions and translocations - rearrangements of the order of the DNA

which can be detected as altering cytological, genetic and physical maps.

CURRENT STATUS OF MAPS AND SEQUENCE INFORMATION IN SEVEN WELL-STUDIED SPECIES.

1. *Escherichia coli* , an enteric bacterium.

Genome size: 4.7 Mb. (Smith et al., 1987; Donna Daniels, personal communication). Approximately 575 kb., or 12% of the genome, has been sequenced and is in the databases (Genbank release 48; roughly March, 1987).

For many years E. coli and its phage have been the principal object of study by molecular biologists, molecular geneticists and biochemists. It is without question the one species which is closest to being completely understood. The genetic map of E. coli is quite extensive. Genetic distances are measured in minutes, being based on the rate of transfer of markers during matings using the F factor. The genome is close to 100 minutes in length; each minute is 40-50 kb. in size.

Progress on the physical map of the *E. coli* genome is good. A complete Not I restriction map is now available (Smith et al., 1987). This restriction enzyme recognizes the sequence GCGGCCGC, which, being 8 nucleotides long, is expected to occur by chance approximately once every 34 kb.. In fact, there are 22 cleavagae sites for this enzyme in the *E. coli* genome (averaging one every 210 kb).

An independent effort to generate a map for all cleavage sites for the enzymes Eco R1, Hind III and Bam H1 is being carried out by Daniels and Blattner at the

University of Wisconsin (Donna Daniels, personal communication). In this study, randomly selected fragments of the E. coli genome cloned in bacteriophage lambda are being analyzed by digestion with those enzymes. Bacteriophage containing overlapping regions can be identified by the similarity of the pattern of these sites and thereby ordered. Groups of phage whose inserts represent overlapping regions of the genome are known as contigs (Staden, 1980; Coulson et al., 1986). Currently, 65% of the genome is covered by 65 contigs, each somewhere in size between 15 kb (the size of a single fragment) and several hundred kb, but averaging about 45 kb.. An estimated 95% of the genome is present in clones which have been analyzed but many clones have not yet been assembled into contigs. At present, over 40% of the genetic map has been assigned to contigs.

This effort differs from the Not I restriction map in the important respect that it generates an ordered set of lambda clones. These clones are immediately useful to anyone wishing to examine DNA corresponding to a gene whose position relative to the map is known. The physical map is correlated with the genetic map at many sites, primarily as a result of the inclusion of clones containing known genes among those examined. The extensive genetic map of E. coli has also been, and will continue to be of use in assembling the physical map. Many clones and/or contigs have been tied to the genetic map by hybridization experiments using preexisting clones of known genes. Clones which are tied to the genetic map can be used to orient the physcial map. The ongoing assembly of a complete map is also aided by the use of large fragments generated by cleavage with Not I and similar enzymes: for example, a purified Not I fragment used as a hybridization probe will identify all of the clones containing portions of DNA from within that fragment: all of these clones will of course be continuous in the genome.

2. *Saccharomyces cerevisiae* , baker's yeast.

Genome size: 14-15 Mb. (Maynard Olson, personal communication; also measured at 2.5 times larger than E. coli by Holliday, 1970). Approximately 350 kb., or 2.4%, of the genome, has been sequenced and is in the databases (Genbank release 48).

An ongoing project to map the *S. cerevisiae* genome is described by Olson et al., 1986. A large fraction of the genome (~95%) is present in clones, which have been joined together into about 400 stretches of contigously mapped genome (contigs). These have been/are being correlated with a now-complete Sfi/Not restriction map for the *S.c.* genome. These combined maps already have significant utility in that it is possible to construct or identify a mapped region of 30-100 kb around virtually any starting point (typically, a cloned gene).

3. *Arabidopsis* , duckweed.

Genome size: 70 Mb. (Elliot Meyerowitz, personal communication). Significant progress on the sequencing of this genome has yet to be made.

This is a rapidly growing plant with a small genome which has been selected by a group of geneticists as a good model organism for the molecular genetic analysis of plants. There are currently two mapping projects under way. One (Goodman, Mass. General) is a cosmid-linking project like that being carried out in nematodes (see below). The second is an RFLP mapping project being carried out by Elliot Meyerowitz at Cal. Tech..

4. *Caenorhabditis elegans*, a nematode worm.

Genome size: 80 Mb. (Wilkins, page 48, cites Sulston and Brenner, 1974). The sequencing of nematode genes is accelerating, but is not significant as of Genbank release 48.

An ongoing project to map the *C. elegans* genome is described in Coulson et al., 1986. The approach is in many ways similar to the yeast project. Like the yeast project it is resulting in a set of clones which cover the genome. One difference is that clones are characterized by a "fingerprinting" method: distances from each cleavage site for one enzyme (Hind III) to the nearest site for a second enzyme (Sau 3A), which cuts more frequently, were measured, and clones sharing a number of such distances were considered to overlap. This makes identification of overlapping regions somewhat easier (as the information is denser) at the expense of not directly yielding information about the restriction map. A second difference between this and the yeast project is that cosmid, rather than bacteriophage lambda clones are being used. Cosmid clones contain larger inserts (30-50, rather than 15-20 kb), but are more difficult to work with, and must be maintained as plasmids in E. coli, a fact which increases the liklihood of recombination (resulting in partially deleted or rearranged inserts). It is not clear to what extent recombination has been a problem with this project. Presently, (as of publication in October 1986) 860 clusters of clones, from 35 to 350 kb in length and totalling about 60% of the genome, have been characterized.

5. *Drosophila melanogaster*, the common fruit fly.

Genome size: 165 Mb. (Spradling and Rubin, 1981). Approximately 265 kb., or 0.16% of the genome, has been sequenced and is in the databases (Genbank release 48).

This organism is the best characterized genetically of any higher (multicellular) eukaryote. The most recent list of extant fly stocks describes 11,597 genetically distinct stocks maintained at various centers throughout the world. The most recent list was published in Drosophila Information Service (DIS), volume 57, June 1982, as "computerized stock list 3". DIS is a longstanding informal publication

which presents unrefereed reports of interest to those engaged in genetic studies on *Drosophila* . Many more stocks and mutants are not in stock centers, but are maintained in the laboratories of interested investigators.

Drosophila is the most complex organism in which it is possible to carry out a saturating screen for mutations of a particular type. A saturating screen is one in which every gene which could mutate to produce the defect being studied has been identified. The availabiltiy of such a technique allows genetic analysis to be thorough in the sense that the entire genome can be examined for the presense of genes which affect a process of interest. The most celebrated example is an exhaustive study of mutations which cause lethality early in development and affect the pattern of the resultant larval cuticle (this study was published as a group of three papers in 1984: Nusslein-Volhard et al., Jürgens et al. and Wieschaus et al.).

The physical mapping of Drosophila is not being undertaken by any single laboratory, but is nevertheless in a very advanced state. There are roughly 500-1000 clones, obtained in various labs in various vectors for various reasons, which have been localized to a map position by in situ hybridization to polytene chromosomes. A listing of these clones, cross-referenced with regard to cytological postition and rearrangements which have breakpoints nearby is maintained by John Merriam and colleagues at the University of California, Los Angeles, and is made available to all *Drosophila* researchers. It was also published in DIS 63, June 1986 (Merriam et al., 1986). Most of these are clones in lambda and have inserts of approximately 15 kb, so 5-10% of the genome (1,000 clones of 15 kb would be 15 Mb, or 9% of the 165 Mb genome) is present in clones which have been roughly ordered by their cytological position. An estimate by John Merriam, announced at the 28th annual National Drosophila Conference, held in Chicago in May of 1987, holds that 10% of the genome is in contigs. These have been established by a variety of investigators who have

been attempting to molecularly clone particular genes by "walking" towards them (a process by which a series of overlapping clones are selected from a genomic library by hybridization with adjacent clones in the contig). These contigs vary in size, ranging from 20 or 30 kb to approximately 1 Mb..

A proposal exists to construct a map of overlapping cosmid clones of the *Drosophila melanogaster* genome (C. Louis, M. Ashburner and H. Jaeckle, submitted to the European Economic Community). This proposal would be much like the nematode project with the important difference that the high-resolution cytological maps available in Drosophila will be exploited in two ways. First, the technique of microdissection cloning, whereby DNA is excised from precise regions of the salivery gland chromosomes and cloned, will be used to generate region-specific clones. These microdissection clones are not of sufficient quality to be used as is, but can be used to sort cosmids by homology into regions before assembling contigs. Later, the position of all contigs with respect to the cytological map will be precisely confirmed by the complementary technique of in situ hybridization whereby cloned DNA is used as a probe in a hybridization to salivery gland chromosomes.

A project recently initiated in the laboratory of Alan Spradling is an attempt to generate of a large number (perhaps 10,000) P transposable element insertions. The P element is a genetic element normally found in some strains but not others. Each of these insertions will be maintained as a strain in which the P element (which is also genetically marked) is present in a single site and causes a mutation. This set could be ordered by in situ hybridization using clone P element sequences as a probe. Such a collection will represent a means of moving back and forth between the genetic and physical maps. It is hoped that ultimately a large fraction of all genes will be tagged by such P element insertions. These will serve as sequences which can be detected by a single hybridization probe (the P element). Once the collection is

established, extremely rapid mapping and cloning of any local area will be possible (on average, one such insertion should be present every 15 kb). Because these insertions will have been previously selected on the basis of having caused a mutation, the correlation of physical and genetic maps will be straightforward.

One of the advantages of Drosophila - that there already exist high quality genetic and cytological maps of the genome - makes data management more complex. It is essential that all genetic and physical mapping information be cross-referenced in some way. Currently this work is being done by John Merriam, through the clone list already described, and by Dan Lindsley, who is overseeing the revision of the classic text Genetic Variations of *Drosophila melanogaster* (1967). This situation is clearly temporary. The funding for Merriam's "Catalogs of cloned DNA, breakpoints and transformed inserts by chromosomal location" is through a temporary grant from the Library of Medicine (under Publication Assistance) and Lindsley plans to halt work after publishing a revised version of the 1968 listing of mutants. Michael Ashburner has expressed some interest in developing a relational database whereby physical and genetic maps will be fully crossreferenced and information on the source of materials (DNA clones and mutant stocks) will be tied to all entries, but this project is currently being considered in theory only.

6. *Mus musculus*, the mouse.

Genome size: 3,000 Mb. (Wilkins, page 150). Approximately 895 kb., or 0.03% of the genome, has been sequenced and is in the databases (Genbank release 48).

This genome is comparable in size to the human genome. However, because genetic analysis of the mouse is well developed, it would be possible to map and sequence DNA from highly inbred strains. This DNA would not be polymorphic, and could include DNA corresponding to any of a number of genes which are responsible for mouse

models of human genetic disease.

7. Homo sapiens , humans.

Genome size: 2,800 Mb. (Gilbert, page 150). Approximately 1,555 kb., or 0.05% of the genome, has been sequenced and is in the databases (Genbank release 48).

On the prospects of a physical map of the human genome, I quote Coulson et al:

"Progress toward a complete nematode map now depends upon the efficiency with which the missing pieces can be found and will increasingly become a communal effort as the "genomic walks" carried out in various C. elegans laboratories are brought together into a common reference data base. Full connection will take a considerable time to achieve and may not be practicable at all. Even near completion, however, the map will be of great value both for studying the large scale organization of the genome and for isolating and characterizing segments that cannot be readily identified in other ways."

If the techniques which are being used on yeast and nematode genomes were to be directly applied to the human genome, one would be left with a thousand gaps to be filled rather than a hundred gaps to be filled. Clearly, new technologies would be needed, particularly technologies for the cloning of large (~Mb range) fragments. These are only now being developed (Burke et al., 1987).

II. Scientific arguments for a concerted effort to map the genomes of other species in parallel with the human genome.

FEASIBILITY AND SPECIAL PROBLEMS IN VARIOUS SPECIES.

1. Proportion of sequence likely to be coding.

Humans, like all mammals, have a genome which consists of at least 90% uninterpretable (and probably meaningless) sequence. This is clear from the sequence and mapping analysis which has been carried out in the vicinity of a

variety of genes. Well-characterized examples include the gene clusters encoding the protein subunits of hemoglobin. In the case of the human beta-globin cluster, the sequence of over 73 kb. of contiguous nucloetides is known (Genbank), and intensive analysis of this region clearly indicates that there is nothing encoded by this stretch of DNA other than five related globin genes whose mRNA products together comprise less than 5% of the total nucleotides (see Orkin and Kazazian, 1984). The abundance of uninterpretable and probably nonfunctional sequence may result simply from the absence of any selective pressure to delete such DNA. Whatever the biological basis may be, the existence of such sequences certainly has practical implications. Determination of the sequence of the entire human genome will involve great expense in determining the order of functionless nucleotides. Stated another way, the large-scale sequencing of DNA from species with fewer nucleotides per gene will yield more information per dollar than the large scale sequencing of the human genome.

2. Unclonable sequences.

There is much concern about the possibility of sequences which are unclonable for some reason. The extent to which this will be a problem is currently hard to estimate, but the feeling of many people is that such sequences will be more common in the human genome, which has a pattern of complex repeats, than in other, smaller genomes.

3. Heterozygosity.

The human population (like <u>wild</u> populations of any species) is highly polymorphic with regard to the sequence of its genes, but the result of a mapping or sequencing project will be a single map or sequence. Furthermore, any source of DNA for this project will necessarily be diploid. Thus, each portion of the human genome will be present twice in the source material, and each region of a map or

sequence will be equally likely to be derived from one copy or the other. In any given region the information contained in the map or sequence could pertain to unusual or even defective genetic information. Such an occurrence would probably escape notice. This problem could be overcome in the case of any gene of interest by the sequencing of DNA from multiple individuals, but would compromise a database.

Of course, this problem could be easily circumvented in the construction of genome maps or sequences from species with genetically well-defined strains (including any of the non-human species described in section I).

4. Correlations with genetic and cytological maps.

For extremely good ethical and pracical reasons, there is no experimental transmission genetics of human beings and it is impossible to systematically identify, in humans, genes involved in a particular process of interest by mutant screens and other appropriate genetic manipualtions. Perhaps the greatest value of a completely sequenced genome will be the ability to rapidly correlate biochemical function (as inferred from the sequence of the gene and its predicted protein sequence) with mutant phenotypes. In the case of the human genome, such correlations would be limited by an inability to generate and genetically analyze mutations. A completely sequenced genome from any of several genetically well-studied species would enable such correlations to be made with greater facility.

5. Effects of size.

A large scale sequencing project would have to involve the independent analysis of pieces which are 10 Mb. or less, with the time spent on any one piece being independent of the size of the total project above that size. Thus, sequencing effort will be roughly proportional to the amount of DNA being sequenced. The mapping of the nonhuman described above would, however, be more efficient because of the greater density of genetic markers and, in the case of *Drosophila*,

polytene chromosomes (a high-resolution cytological map). The total size of the combined *E. coli*, *S. cerevisiae*, *C. elegans* and *D. melanogaster* genomes is 265 Mb., while the size of the human genome is 2,800 Mb.. Thus, a large scale sequencing project which was restricted to the human genome would be only 9% less total effort than a project to completely sequence the genomes of *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens* . Such a restriction of the project seems particularly ill-advised when the benefits of this additional sequencing are considered (see below).

RELATIVE VALUES OF HUMAN AND NONHUMAN MAPS OR SEQUENCES

It may be naive to assume that the only sequence which will be of use in the improvement of human health and welfare is the human sequence. There follows a discussion of the great value of a pre-existing database from a genetically tractable organism for the interpretation and understanding of human data.

The human genome is a complex set of instructions for the construction of a human being, and we would all very much like to see those instructions and understand them. This would involve knowing two things. The first is the sequence and the second is the means by which the sequence specifies the construction of a human being. It is with regard to the second of these concerns that the fruit fly genome has some very attractive features. The fly genome is a set of instructions for the construction of a fly which

1) uses a very similar language to that used by the human genome,

2) is shorter by a factor of about 20, and

3) can be experimentally manipulated with much greater facility.

By a similar language I mean both that the signals for the expression of genes (such as those for transcriptional regulation and RNA processing) are similar and

that many of the mechanistic details of development itself are conserved. In fact, the majority (if not all) of the general classes of molecule which have been identified as functioning in mammalian development can be found in Drosophila. - This includes cytoskeletal components (proteins involved in the structure of cells), metabolic enzymes, many specific types of growth factors and their receptors, steroid hormones and steroid hormone receptors, and even the insulin receptor. Conversely, mammalian counterparts to molecules such as the products of the homeotic genes (which have been shown genetically to act in the fly in the control of such developmental decisions as whether to make an antenna or a leg) have been found in mammals, although their role there remains unestablished.

The fact that the fly genome is only 5% the size of the mammalian genome means that molecular analysis will be easier. The entire genome could be mapped or sequenced with less than 5% of the effort required to sequence the human genome (genome size as a factor in mapping and sequencing projects is discussed in the preceding paragraphs).

The third property of the fruit fly mentioned above is the most significant. If we wish to understand the functioning of a gene the most useful experiment is to carefully examine the phenotype of mutations in that gene. Further information can be obtained by identifying genes which interact with the gene in question - once again, the techniques which allow this are genetic techniques available in Drosophila (and in yeast and in nematodes) but not in humans. In this way the function of a gene can be understood in terms of its protein product and its interactions with the products of other genes. Armed with such an understanding of Drosophila genes, it will then be possible for investigators to fully understand the human genetic disease in terms of mutations in the corresponding human genes (see Figure 3). Without this approach, there may not be a way of obtaining similar

information about the gene products and the proteins with which they interact.

An example will serve to illustrate how this can work. Oncogenes are genes whose mis-expression (usually as a result of mutation or incorporation by a retrovirus) can lead to the transformation of normal cells into cancerous cells; both overexpression of the normal protein and expression of an altered protein product have been observed to cause the cancerous state. Many oncogenes have been identified in mammalian systems by a variety of techniques, but in no case has the normal cellular function of such a gene been clearly identified. However, there are now three cases of genes in Drosophila, previously identified by mutational analysis, which bear considerable homology to mammalian oncogenes. One is the gene *sevenless*, which is essential for the formation of the R7 photoreceptor cell in the eye of the fly (Hafen et al., 1987). In the absense the function of the *sevenless* gene this particular cell type fails to form and is replaced by another. A second case is that of the gene *dorsal*, which must be expressed in the mother for normal embryonic development to ensue (Ruth Steward, personal communication - data was presented at the 1987 National Drosophila Conference). A third is the gene *wingless*, which is necessary for the normal development of segments in the early embryo (Nick Baker, personal communication - data was presented at the 1987 National Drosophila Conference). All of these genes are quite homologous to mammalian oncogenes, and in all of these cases there exist other genes which, by various genetic criteria, are known to encoded products which interact with the oncogene homolog. Thus, it will now be possible (by cloning these interacting genes and analyzing their products) to determine the biochemical mechanisms by which oncogene homologs function in Drosophila development. Once that has been accomplished, the overall understanding of the action of oncogenes will have been greatly advanced in a way which would have been impossible without genetics.

The establishment of a complete map of the human genome, to the nucleotide, would be of value in the diagnosis and understanding of human genetic diseases. It will allow guesses as to the nature of diseases which will then have to be confirmed by experimentation. "The" human sequence, useful as a database, will have to be complemented by the sequence of affected individuals. The sequence will not provide any simple means of identifying those genes which are the basis of an inherited disease; careful genetic mapping through the detailed studies of affected families will still be necessary.

The human genome sequence will be of value as a complete database of genetic information from a single organism. Benefits would follow from the resultant greater understanding of biology and biochemistry. A complete sequence will be especially valuable as an indicator of the function of a particular gene or protein, to be confirmed by the techniques of molecular biology and biochemistry. This kind of process will only begin to operate as the sequence reaches completeness, when it becomes possible to usefully search the sequence for a match to a given protein or nucleotide sequence, and thereby identify all potential genes encoding proteins of a given type.

Not all of the benefits to be derived from the complete sequence of the human geneome depend on its being the human genome. In fact, there are many reasons why *Homo sapiens* is a poor choice for this endeavor (these are discussed above and include genome size, lack of experimental transmission genetics, etc.). Once the human sequence becomes technically feasible, it will also be feasible to rapidly determine the sequence around the location of a gene causing human disease or the sequence of the human counterpart to a gene whose function has been established in experimentally tractable species (see Figure 3).

ASSOCIATED TECHNOLOGIES WHICH WILL INCREASE THE USEFULNESS OF THE MAP OR SEQUENCE.

The sequence itself will not reveal the location of genes which have been identified genetically (as opposed to those which have been identified by knowledge of their protein or RNA products). This class of genes includes those which cause human disease. Thus, the current efforts directed towards the cloning of genes will not be eliminated by a complete sequence; rather effort currently expended towards finding a clone for a particular gene will be redirected towards localizing that gene within the sequnence.

To identify genes within known sequence, one typically looks for regions which could code for protein and then predicts the protein product by making informed guesses about the sites at which translation is initiated and terminated. The biggest problem here is probably that of correctly evaluating the pattern of splicing which fuses coding exons.

To quote from Weston and Barrell,1986:

"The wide variation in the sequences of functional splice donor and acceptor sites makes it extremely inadvisable to attempt any specific specific predictions of splicing from DNA sequence alone."

This was written in a paper from the MRC Laboratory of Molecular Biology, the location of some of the greatest expertise in the analysis of sequence information (I refer in particular to the programs of Staden).

A recent state-of-the-art method for the prediction of splice sites (Nakata et al., 1985) reports an accuracy of greater than 80% correct splice site prediction. The quality of methods for this kind of prediction is limited by the degree to which biological signals such as splice sites conform to standard rules, and it appears likely that the accuracy of such methods will not improve greatly until or unless there is a

major breakthrough in the understanding of splice site choice. With this accuracy, the correct protein sequence could be deduced from DNA sequence alone less than 35% of the time in a gene with 5 exons (ten splice sites). Thus, it will generally be possible to locate pieces of protein-coding information within a large sequence, but it will not be possible to determine complete protein sequences with anything other than a very modest degree of success.

Genes, particularly their protein-coding regions, are more conserved than are surrounding sequence. This fact has been successfully used to identify coding regions in the study of several Drosophila genes. Comparisons were made between the nucleotide sequence of the same gene in different species. In the case of the engrailed gene a comparison was made between two species within the genus Drosophila (Kassis et al., 1986), and, in the case of the scarlet gene, between Drosophila and the blow fly (Tony Howells, personal communication, presented at the 1986 National Drosophila conference at Asilomar, California). In both of these cases protein-coding regions were clearly identifiable as stretches of nucleotides with the greatest level of conservation. A similar approach would be made possible by the existence of a complete sequence from any two higher (multicellular) eukaryotic species.

A word of caution comes from a study which describes a portion of the Drosophila genome in which there exists a region of slowly-evolving DNA which apparently does not encode a protein (Marten and Meyerowitz, 1983)

Comparisons between species with different degrees of relatedness will be useful in the analysis of different sets of genetic information, due to differing rates of variation among different proteins and differing rates of sequence divergence in portions of a gene (e.g. protein-coding regions, introns, promoter regions, etc.).

With regard to the human sequence, a comparison with mouse sequence would

probably be most useful in the identification of protein coding regions, a comparison with the Drosophila or nematode sequence will be most useful in the identification of sequences encoding proteins with functions essential for all multicellular organisms, and a comparison with the yeast sequence will be most useful in the identification of sequences encoding proteins with functions essential to the life of a eukaryotic cell.

III. Recommendations based on the above arguments.

Recommendation 1: Initial efforts should be directed largely, if not exclusively, towards further improvements in the necessary technology.

First, let us first consider the expense of sequencing.

The current cost is roughly $1.00 per nucleotide (Wada, 1987; Hood and Smith, 1987). At this rate, the cost of sequencing the human genome would be three billion dollars. Cheaper technologies are expected to be developed soon (Wada, 1987; Hood and Smith, 1987), but are as yet unproven.

A reliable number for the rate, in kb per man-year, of sequencing with current state of the art technology, is difficult to come by. The longest complete published sequence, that of Epstein Barr Virus, is roughly 172 kb and consumed approximately 15 man-years (The rate is therefore 10 kb/man-year). A more recent figure is approximately 50 kb in about two years time (Weston and Barrell, 1986 - the short unique region of cytomeglovirus), or 25 kb/ man-year. In discussions with G. Hatfull, who participated in the sequencing of Epstein Barr Virus sequencing, I was given 50 kb/man-year as a realistic value for an industrius individual. If that number were 60 kb/ man-year, then the human genome would represent an investment of 50,000 man-years. 10,000 man-years would be a rate of 300,000 kb/man-year, a rate which may be conceivable in the not-too-distant future, but which is beyond anything which is currently being done anywhere, even with the most current state-of-the-art technology.

Another way of doing the calculation is to use the number of nucleotides of raw sequence which can be readily generated by one individual in a very busy day (250 nucleotides of reliable sequence on each of 40 clones: 10,000 nucleotides) and the level of redundancy required for an acceptable error rate with a random-clone approach (~10), which suggests that good sequence can be generated at a rate of 1 kb/day.

(1 kb/day)(250 day/ manyear)(1 genome/3 x 10^6 kb) = 1 genome/12,000 man-years.

That counts only time spent doing sequencing reactions and running gels, and is very optimistic about the stamina of the worker, so it clearly represents a lower limit on the number of manyears required for the sequence.

Thus, any optimistic but reasonable estimate for the time required for the sequencing alone (using current state-of-the-art technology) would fall between 10,000 and 40,000 man years.

Second, it is necessary to consider the time required for the mapping, cloning and data management that would be necessary. These steps would probably have the greatest impact at the level of the entire project, particularly in the generation of a map and the cloned DNA to be sequenced. An estimate of the effort required for mapping the human genome with current state-of-the-art technology is impossible because the technology appropriate for the task is only now emerging. Thus, estimates of the time required to sequence the human genome with current methods are meaningless, but speculative estimates based on extrapolation of current efforts to map much smaller genomes are possible. The nematode and yeast genomes are on the order of 0.5 to 3% the size of the human genome, and have currently consumed 5-10 man-years each to reach an uncertain level of completion. The greater complexities of the human genome, particularly the problem of repeated sequences

(which may or may not be disposed in a way which makes mapping difficult), further adds to the uncertainty. However, putting these numbers together and making a wild stab, an estimate of between 2,000 and 10,000 man-years for mapping and cloning seems reasonable.

A similar degree of uncertainty surrounds the computational and data management aspects. However, it seems likely that, although the job is enormous, it will require more money than time.

Thus, the sequencing itself will probably be a much larger job than the mapping, cloning and data management, with the total investment of effort being in the range of 12,000 to 50,000 man-years. Issues such as the availability of sufficient numbers of properly trained individuals will need to be addressed further.

In short, a full scale mapping and sequencing project would be an enormous expense if current methods were used. Further improvements in technology are essential if this project is to be considered realistically.

Recommendation 2: Competitive invesigator-initiated grants should make up a large portion of the funding.

Whatever form a genome mapping or genome sequencing initiative takes, the most efficient and reliable method may still be the small-science approach, with investigator-initiated proposals, initially having to do with improvements in methodology, later with mapping, and eventually (if it is still judged to be worth the cost) sequencing itself. The mapping and sequencing of genomes is a large undertaking, which obviously needs centralized large-scale data managment. However, it is not necessary that the work itself be centralized. In conversations with people who are working on related technologies one is lead to the conclusion that no current method of mapping and sequencing is clearly superior to others. Nor is it

clear what kind of map is desirable. However, many good ideas exist, and it is likely that the best ideas will continue to come from those directly involved in the research. Centralization might lead to a homogeneity of approach which would be stifling. Conversely, competitive investigator-initiated grants of the sort which have funded the bulk of biomedical research over the last ten years would probably serve to bring the best ideas forward.

Recommendation 3: Policy-makers should excercise vigilence to insure that related basic science research is not compromised as a result of a genome initiative.

There is very much that the sequence of the human genome will not tell us. I quote from the draft of the OTA panel report:

"We have very limited knowledge of how the expression of genes is controlled. What sequences of the DNA turn genes on and off at the right time for correct development and differentiation? We do not understand how the coordinated control of genes is accomplished. We expect that vital elements that exist in the genome have not even been imagined. The human genome has been called the book of man; it contains the instructions that describe each human. It is time to obtain a copy of the book to begin to understand what the text means."

As described above, the answers to these questions will not be apparent from the sequence itself. They will only come from experimental approaches involving manipulation of the DNA and study of the effects of that manipulation. Such research is moving rapidly ahead at the present time. There is a danger that this research, essential for the interpretation of the sequence, will be regarded incorrectly as redundant or unnecessary as the sequence becomes available.

Also, it is likely that ongoing research aimed at solving medical and biological

problems will, by itself, lead to much of the scientific progress derivable from the human genome initiative. Important portions of the sequence will be identified and obtained in the same manner as they have been, and this will occur at an accelerating rate as technology improves. It appears that with equivalent overall funding, scientific progress of the sort which would result from the genome initiative would most likely be greatest without the genome initiative! Only if entirely separate funds are made available could the genome initiative proceed without slowing progress towards the goals it is meant to serve.

Recommendation 4: Improvements in the handling of nucleotide sequence and genetic information should be given priority.

Centralization may be necessary for the coordination of genetic map and sequence information. Currently, nucleotide sequence databases run well behind (up to and beyond one year, depending on a variety of factors) and do not offer a variety of imaginable and useful services (such as the translation of nucleotide sequences into protein sequences). Even without a genome initiative, improvement in repositories of nucleotide sequence and genetic information would be desirable.

Recommendation 5: The genomes of several other genetically tractable species should be physically mapped and/or sequenced in parallel with the human genome.

The goals inherent in a large scale effort to sequence the human genome may be better served by a broader effort directed more efficiently at the benefits to be achieved. I would recommend the establishment of an program to sponser and coordinate the generation and processing of genetic mapping and sequencing data from several species.

A project to completely sequence a smaller genome would be useful for the development and testing of new techniques, without the danger of locking the much larger human project into an inferior technology. I'm thinking - especially of software for the overall organization of sequence information. In this scheme the project could graduate from smaller to larger genomes, having learned valuable lessons and ready for another generation of answers (compare the way the Sanger lab sequenced first ØX176, then lambda, then Epstein-Barr virus).

Given the arguments elaborated in earlier sections, it seems fairly indisputable that the sequence of the simpler genome of a genetically tractable organism such as *S. cerevisiae* or *D. melanogaster* would be easier to come by, would contain more information (the sequence of more genes), and would be more easily interpreted than any similarly sized region of the human genome. Thus, if a certain amount of money (say $165 million) were spent to sequence a certain number of nucleotides (say 165 million), then more value would be obtained for that money from the complete Drosophila genome than from 5% of the human genome.

One central agency might also take over sequence database services for all species, would support independent research on computational and statistical aspects of nucleotide sequence analysis, and would support independent research on sequencing and mapping techniques. Eventually, it might be worthwhile to complete the sequence of the human genome, but only once its time had come.

Figure Legends.

Figure 1. An example of genetic linkage, using three hypothetical genes. A and B are on different chromosomes, and A segregates away from B during meiosis 50% of the time (no linkage). In contrast, gene C is present on the same chromosome as A, and alleles for A and C cosegregate unless recombination occurs between them. Recombination occurs between A and C 10% of the time, so that linkage appears to be incomplete. It is the frequency of recombination between different genes on the same chromosome which allows the construction of genetic maps.

Figure 2. An example of salivary gland chromosomes. This photomicrograph shows the banded pattern of a salivary gland chromosome as seen under phase contrast light microscopy. The pattern of banding is characteristic of the particular chromosomal region, and allows individual genes to be localized on the cytologically visible chromosomes with a resolution of about 10 kb..

Figure 3. A flow chart for the application of information obtained in experimentally advantageous species. This scheme illustrates how knowledge of which genes interact in the development of an organism such as the fruit fly can be used to identify genes of importance in the human genome.

References.

Burke, D.T., G. F. Carle and M.V. Olson (1987) Cloning large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 236:806-812.

Coulson, A., J. Sulston, S. Brenner, and J. Karn (1986) Toward a physical map of the genome of the nematode *Caenorhabditis elegans* . Proc. Natl. Acad. Sci. U.S.A. 83:7821-7825.

Fan, J.B., C.L. Smith and C.R. Cantor (1987) Construction of a physical map of the *Schizosaccaromyces Pombe* genome. Genetics 116:s4.

Funanage, V.L. and Myoda, T.T. (1986) Localization of Chinese Hamster Dihydrofolate Reductase Gene to Band p23 of Chromosome 2. Somatic Cell and Molecular Genetics 12:649-655.

Gilbert, S.F. (1985) Developmental Biology. Sinauer Associates, Inc..

Goodenough, U. (1984) Genetics. CBS College Publishing.

Hafen, E., K. Basler, J.-E. Edstrom and G. M. Rubin (1987) *Sevenless* , a cell-specific homeotic gene of Drosophila, encodes a putative tranmembrane domain receptor with a tyrosine kinase domain. Science 236:55-63.

Holliday, R. (1970) Symp. Soc. Gen. Microbiol. 20: 362

Hood, L. and L. Smith (1987) Genome sequencing: how to proceed. Issues/Sci. & Tech.

Jürgens, G., E. Wieschaus, C. Nüsslein-Volhard, and H. Kluding (1984) Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster* II. Zygotic loci on the third chromosome. Roux's Arch Dev Biol 193:283-295.

Kassis, J.A., S.J. Poole, D.K. Wright and P.H. O'Farrell (1986) Sequence conservation in the protein coding and intron regions of the *engrailed* transcription unit. EMBO J. 5:3583-3589.

Lindsley, D.L. and E.H. Grell. (1968) Genetic Variations of Drosophila melanogaster. Carnegie Institution of Washington

Martin, C. and Meyerowitz, E.M. (1986) Characterization of the boundaries between adjacent rapidly and slowly evolving genomic regions in *Drosophila* . Proc. Natl. Acad. Sci. U.S.A. 83:8654-8658.

Merriam, J., S. Smalley, A. Merriam and B. Dawson. (1986) The Molecular Genome of *Drosophila melanogaster* : Catalogs of cloned DNA, breakpoints and transformed inserts by chromosomal location. Drosophila Information Service 63:173-264.

Nakata, K., M. Kanehisa and C. DeLisi (1985) Prediction of splice junctions in mRNA sequences. Nucleic Acids Res. 13:5327-5340.

Nüsslein-Volhard, C., E. Wieschaus and H. Kluding (1984) Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster* I. Zygotic loci on the second

chromosome. Roux's Arch Dev Biol 193:267-282.

Olson, M.V., J.E. Dutchik, M.Y. Graham, G.M. Brodeur, C. Helms, M. Frank, M. MacCollin, R. Scheinman and T. Frank (1986) Random-clone strategy for genomic restriction mapping in yeast. Proc. Natl. Acad. Sci. U.S.A. 83:7826-7830.

Orkin and Kazazian (1984) The mutation and polymorphism of the human beta-globin gene and its surrounding DNA. Ann. Rev. of Genet. 18:131-171.

Smith, C.L., J.G. Econome, A. Schutt, S. Kleo and C.R. Cantor (1987) A physical map of the *Escherichia coli K12* genome. Science 236:1448-1453.

Staden, R. (1980) Nucleic Acids Res. 8: 3673-3694.

Spradling, A.C. and Rubin, G.M. (1981) Drosophila genome organization: conserved and dynamic aspects. Ann. Rev. Genet. 15:219-264.

Sulston, J. and S. Brenner (1974) The DNA of *Caenorhabditis elegans*. Genetics 77:95-104.

Wada, A. (1987) Automated high-speed DNA sequencing. Nature 325:771-772.

Weston, K. and B.G. Barrell (1986) Sequence of the short unique region, short repeats and part of the long repeats of human cytomegalovirus. J. Mol. Biol. 192:177-208.

Wieschaus, E, C. Nüsslein-Volhard, and G. Jürgens (1984) Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster* III. Zygotic loci on the X-chromosome and fourth chromosome. Roux's Arch Dev Biol 193:296-307.

Wilkins, A.S. (1986) Genetic Analysis of Animal Development. (John Wiley & Sons, New York).

Telephone consultants.

Burks, Christian. Los Alamos National Laboratory. Santa Fe, NM (505) 667-6683

Church, George.

Daniels, Donna. Blattner lab. University of Wisconsin, Madison. (608) 262-2534

Hatfull, Graham. Yale University, New Haven, CT. (203) 785-4584

Lane, Michael. (315) 473-5855

Merriam, John. University of California, Los Angeles. (213) 825-2256

Meyerowitz, Elliott. Cal. Tech., Pasadena, CA. (818) 356-6895

Olson, Maynard. Washington University, Saint Loius, MO (314) 362-2721

Smith, Cassandra. Columbia University, Med.. New York, NY (212) 305-4145

Smith, Lloyd. Cal. Tech., Pasadena, CA. (818) 356-4901

Spradling, Alan. Carnegie Institution of Washington, Baltimore, MD (301) 467-1414

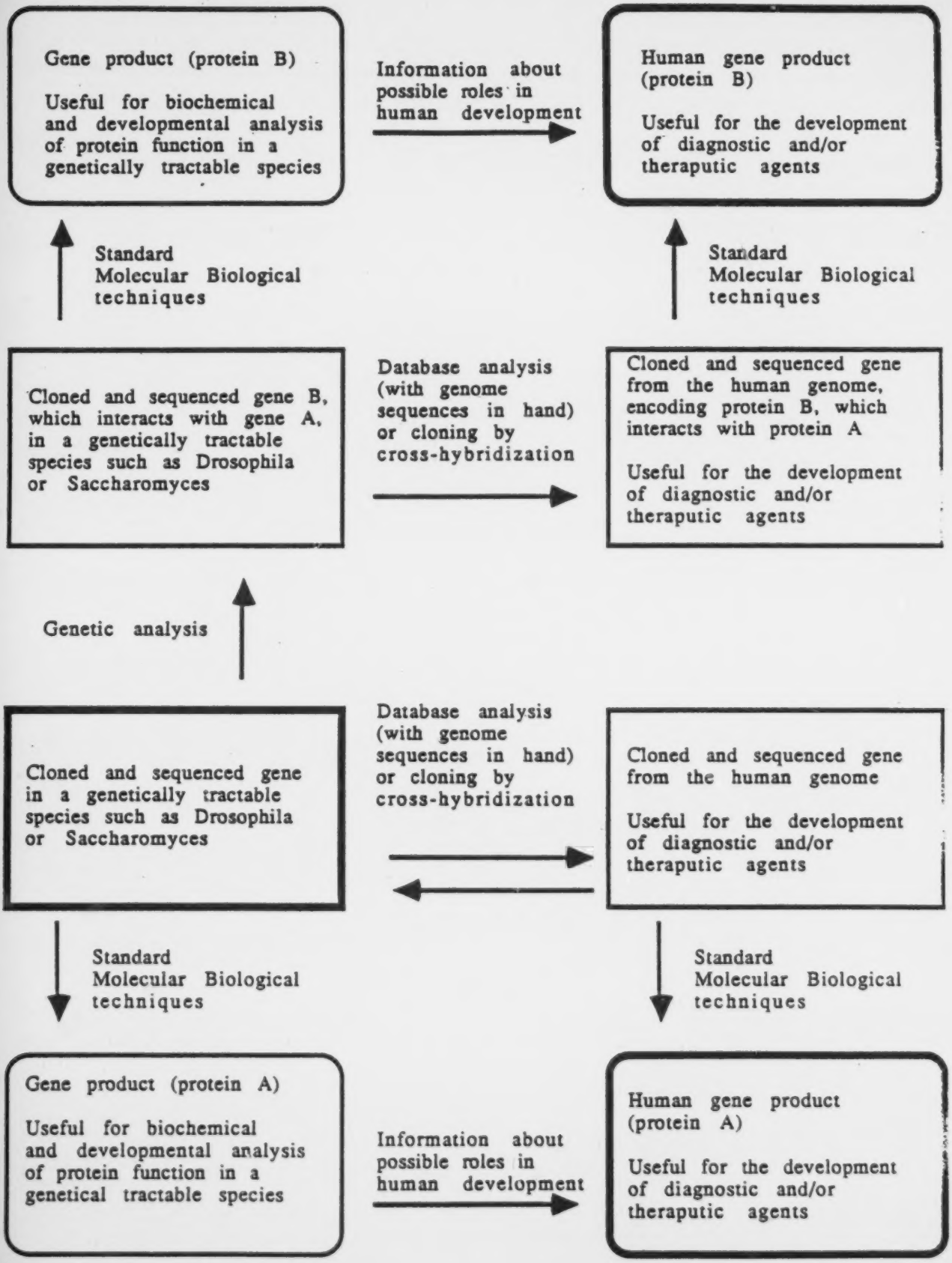
Parental genotype:

A C	B
a c	b

Genes transmitted to progeny:

25 % A B	but	45 % A C
25 % A b		5 % A c
25 % a B		5 % a C
25 % a b		45 % a c

Fig. 1


Gene product (protein B)
Useful for biochemical and developmental analysis of protein function in a genetically tractable species
Information about possible roles in human development
Human gene product (protein B)
Useful for the development of diagnostic and/or theraputic agents
Standard Molecular Biological techniques
Standard Molecular Biological techniques
Cloned and sequenced gene B, which interacts with gene A, in a genetically tractable species such as Drosophila or Saccharomyces
Database analysis (with genome sequences in hand) or cloning by cross-hybridization
Cloned and sequenced gene from the human genome, encoding protein B, which interacts with protein A
Useful for the development of diagnostic and/or theraputic agents
Genetic analysis
Cloned and sequenced gene in a genetically tractable species such as Drosophila or Saccharomyces
Database analysis (with genome sequences in hand) or cloning by cross-hybridization
Cloned and sequenced gene from the human genome
Useful for the development of diagnostic and/or theraputic agents
Standard Molecular Biological techniques
Standard Molecular Biological techniques
Gene product (protein A)
Useful for biochemical and developmental analysis of protein function in a genetical tractable species
Information about possible roles in human development
Human gene product (protein A)
Useful for the development of diagnostic and/or theraputic agents

MAPPING THE HUMAN GENOME: Experimental Approaches for Cloning and Ordering DNA Fragments

Richard M. Myers, Ph.D.
Department of Physiology, S—762
University of California
513 Parnassus St.
San Francisco, CA 94143

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

INTRODUCTION

The purpose of this report is to discuss experimental methodologies that can be used to construct a physical map of the human genome. In this report, the physical map is is defined as the alignment of all the DNA fragments in our chromosomes, correctly ordered and marked, on average, every 500 base pairs by restriction enzyme cleavage sites. Construction of a physical map of any region of DNA involves fragmenting that DNA by restriction enzymes, cloning all the fragments into some sort of vector to establish a library, and then determining the order of each of the fragments and the restriction enzyme cleavage sites within them as they appear in the chromosome. For the human genome, which is composed of about three billion base pairs (Vogel and Motulsky, 1986), over six million restriction enzyme cleavage sites need to be mapped in order to construct a complete physical map.

In addition to a physical map, it is also possible to construct another type of chromosomal map, called a genetic map, for the genome of any organism from which specific genes can be identified. A genetic map of a genome is a description of the relative locations of genes along the chromosomes. In this case, "genes" are defined "genetically", which means that they are identified by specific observable changes in the phenotype of the organism when they are altered by mutation. The way that a phenotype is inherited in the offspring of specific matings provides information about the nature of the gene that specifies the phenotype, such as its dominance or recessiveness. In the case of humans, many single genes have been identified that result in disease when mutated; these genes are the main data that have been used to construct the existing human genetic map. The relative locations of genes are determined by the frequency with which they recombine with one another during sexual conjugation or meiosis; genes close together will recombine infrequently, whereas genes far apart will recombine more frequently.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Although a genetic map constructed on the basis of recombination frequencies allows the relative order of genes to be determined, it does not accurately indicate the physical distance, in terms of base pairs, between genes on a chromosome. This inaccuracy results from the unequal rates of crossing over that occur along the lengths of each chromosome. However, a genetic map, even a partial one, is useful to have during construction of a physical map, since it allows the assignment of known genes to some regions that are being physically mapped. In the case of constructing a physical map of the human genome, knowledge of the existing genetic map can be used to initiate the project so that the first information gathered will be at regions known to contain functional genes that are important in human disease. Appendix 2 contains a description of the techniques used in genetic mapping and some discussion of the ways that the human genetic map can help in constructing the physical map and vice versa.

A physical map of any genome would be extremely useful in its own right, providing researchers with enormously valuable information about that organism. In the case of humans, not only would a physical map teach us much about the organization of genes and other DNA sequences along the lengths of our chromosomes, but it would also provide us with extremely powerful tools to diagnose genetic diseases and to isolate the genes responsible for these diseases. The isolated genes will help us to begin to understand the causes of genetic defects and ultimately to treat them. As mentioned above, during its construction, bits and pieces of the physical map can be aligned with our existing genetic map, providing information about the physical locations of genes that have been identified genetically; this information will help in disease diagnosis, in understanding genetic recombination in mammalian cells, and in expanding the available genetic map itself. A physical map and its material side-products (in the form of DNA clones) would also be extremely useful, probably even necessary, tools to have in hand before an organized effort is mounted to determine the nucleotide sequence of the human genome. Although the sequence (which is the ultimate detailed physical map)

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

could in principle be determined by sequencing random fragments of the genome, a much less laborious approach would be to clone and physically map all the fragments in the genome, and then to begin sequencing those fragments in an ordered fashion. It is very likely that in the course of mapping the human genome, we will benefit from as yet unimagined discoveries and advances in technology. Unquestionably, the construction of a physical map of the human genome will lead to immeasurable gains in our understanding of biological processes.

Although there is reason to be enthusiastic about the enormous benefits that will certainly come from determining a physical map of the human genome, there is also cause for skepticism about charging headlong into such a project without careful thought about its costs and without serious considerations of the experimental difficulties that will be encountered. Several estimates of the dollar costs for the mapping project and the sequencing project have been made (Lewin, 1986a, 1986b, 1987; Smith, 1986; Sinsheimer, 1986; Gall, 1986); it is very difficult to evaluate these estimates this report will not attempt to do so. Most scientists feel very strongly that the mapping and sequencing projects should not be done at the expense of any other type of biological and medical research. A tremendous amount of our understanding of biological phenomena and many technical advances have come from serindipidous discoveries, collaboration between researchers in different disciplines, and from basic or applied researchers being able to make connections between seemingly unrelated phenomena. It would be a terrible mistake to destroy this way of doing science for the sake of a large Manhattan-style project to map and sequence the human genome.

In addition to the actual costs of mapping the genome, we must also consider the experimental difficulties that will be encountered in attempting to obtain a complete map. There is no question that with current technologies and a large organized effort it would be possible to map most (75 to 90 percent) of the human genome. However,

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

obtaining the rest of the map will be very difficult, perhaps beyond our ability to estimate. This difficulty stems partly from the already-recognized phenomenon of "unclonable" or "difficult-to-clone" sequences present in our genomes. We may be surprised to find that some repetitive sequences, which are rampant in the human genome, are not stable when cloned into bacteria. A second problem is a statistical one that arises in any attempt to obtain every restriction fragment in a genome as a bacterial clone; as we get closer and closer to "saturation", it becomes more and more difficult to isolate a fragment that is missing in the map. So after an initial burst of obtaining many cloned DNA fragments that can be linked up physically on the map, it becomes difficult to fill in the gaps between clones. Based simply on statistical arguments and ignoring the problems with non-randomness of cloning, it is estimated that a five-fold redundancy in mapping experiments will be necessary to obtain a 99 percent complete physical map (Clarke and Carbon, 1976; Maniatis et al., 1982; see Appendix 3). These experimental difficulties are not insurmountable, and may be minimized by new technical advances; however, it is important that they are sufficiently addressed in the current debates about whether a full-scale effort to map and sequence the human genome should be undertaken.

The remainder of this report is concerned with the experimental strategies that could be used to construct a complete or near complete physical map of the human genome. First, the overall experimental approach is discussed. A detailed description of each experimental step will follow.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

EXPERIMENTAL STRATEGY

Before discussing the various experimental methods that will be employed in the human genome mapping project, a few technical points, definitions, and some numbers need to be discussed. In addition to the following definitions of terms, a glossary appears in Appendix 1. First, the genome mapping project will require the isolation, via molecular cloning techniques, of every human DNA fragment in the genome several times over, in the form of DNA inserts cloned into some type of DNA vector. Several different types of vectors have been developed extensively since the beginnings of recombinant DNA technology in the early 1970's. Plasmid vectors are small (1,000-10,000) bp circular DNA molecules that can carry foreign DNA inserts up to 12,000 bp; efficiencies of cloning of 10^6 clones/ug insert DNA can be achieved with plasmid vectors (Bolivar and Backman, 1979; Bernard and Helinski, 1980). Bacteriophage lambda vectors are about 20,000 bp and can accept foreign DNA molecules up to about 23,000 bp; cloning efficiencies are 10^8 clones/ug insert DNA (Williams and Blattner, 1980; Hendrix et al., 1982). Cosmid vectors are small (5,000 bp) plasmids, which are derivatives of lambda, that can carry inserts up to about 45,000 bp; the efficiency of cloning in cosmids is about the same as that for lambda vectors (Hohn and Collins, 1980; Ish-Horowicz and Burke, 1981; Evans and Wahl, 1987). Other bacteriophage genomes have been used as vectors (ref), with the aim of increasing maximum insert size and efficiency of cloning; however, none of these other systems have been developed to the point of surpassing the three E. coli vectors described here. The extreme bias of researchers towards E. coli plasmid, cosmid, and lambda vectors is due to the extensive knowledge that we have of the genetics and life histories of these particular organisms (Herskowitz, 1973; Williams and Blattner, 1980; Hendrix et al., 1982; Glass, 1982). Because both insert size and efficiency of cloning is highest in cosmid vectors, it is almost certain that they will be the vector of choice for the genome mapping project.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Thus, if this approach is followed, at the completion of the physical mapping project, a bank of cosmid clones that represent the entire human genome will be in hand; each of these cloned DNA fragments, which are about 40,000 bp in length, will be lined up in the order that they appear in human chromosomes. The positions of cleavage sites for two different restriction enzymes within each of these cosmid inserts will be known. With these materials and information, then, a map of the entire genome could be drawn from end to end, with "landmarks", in the form of restriction enzyme cleavage sites, spaced an average of about 500 bp or so apart. The end result of the physical mapping project is a high-resolution restriction enzyme cleavage map of the entire human genome.

There are three general approaches that have been imagined as the best ways to approach a massive project to map the human genome. One approach would be to begin immediately to make an ultimate high-resolution map, which is the nucleotide sequence of the entire genome. However, random sequencing of cloned DNA fragments representing the entire human genome would result in a high degree of redundancy of sequencing, so that most DNA fragments would be sequenced many times before the project was complete. With current technology (which includes some automation), this approach would be prohibitively expensive; great strides in technology will need to be made before a random sequencing scenario can be seriously considered. A second, less ambitious approach to map the human genome would be to clone the genome in overlapping fragments into a cosmid vector, followed by ordering all of the fragments and determining their restriction maps. Because of the difficulty in ordering such a large number of fragments from one large pool of data, however, it is generally agreed that a third approach is the best strategy: Divide and subdivide the genome into smaller and smaller segments before beginning the restriction mapping itself. After subdivision, restriction maps of these smaller segments are determined, and the information is linked together to form continuous maps of whole chromosomes. In principle, this strategy

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

could be broken down into the following five steps:

1. The isolation of each human chromosome.

2. The division of each chromosome into a collection of overlapping DNA fragments 0.5 to 5 million bp in length.

3. The subdivision and isolation of each of these chromosomal fragments into overlapping DNA fragments 40,000 bp in length.

4. The determination of the order of the 40,000 bp DNA fragments as they appear in the chromosomes and the determination of the positions of recognition sites for a frequent-cutter restriction enzyme within each of these 40,000 bp fragments.

5. The use of the mapping information gained in Step 4 to link together each of the overlapping 0.5 to 5 million bp fragments isolated in Step 2.

A flow diagram depicting these steps is shown in Figure 1. In this scheme, the first two steps are important because they allow the most difficult steps, ordering the 40,000 bp fragments and determining their restriction maps, to be performed on a small subset of the whole genome. As discussed below, most ordering and mapping strategies would be extremely difficult to execute with a pool of fragments representing an entire genome, particularly one as large as the human genome. The strategy described here would involve ordering and mapping the 40,000 bp fragments in each of the overlapping 0.5-5 million bp subfragments, and then using the information in the overlaps to connect each of the subfragments to form a map of a whole chromosome.

These experimental steps are discussed in detail below. It is possible to accomplish the goals of each step by more than one experimental approach. Some alternative approaches and their relative merits will be presented.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Step 1: Isolation of each human chromosome.

Two methods can be used to obtain single human chromosomes separated from all the others. The first, which has been used for many years, involves the formation of hybrid cell lines carrying a single human chromosome in a rodent cell. The second is the physical purification of intact chromosomes by flow cytometry.

In a sense, part of the first step of this project has been done. The collective works of many researchers over the past two decades has resulted in somatic cell hybrids carrying several of the human chromosomes in separate cell lines (Shay, 1982). Thus, rodent cell lines carrying an intact human chromosome 11, and no other human DNA, are available. Similarly, lines are available for several of the remaining autosomes and for the X chromosome. The cell lines are generally constructed by inducing the fusion of rodent and human cells by polyethylene glycol or other agents, followed by limited dilution of the cells and selection for retention of a specific human chromosome on the basis of specific biochemical markers. For example, cultured mouse cells carrying a mutation in the gene for coding thymidine kinase can be used in cell hybrid experiments to select for a new cell carrying human chromosome 17, on which the human thymidine kinase gene is located (Boone et al., 1972). After cell fusion and selection, the resulting cells are single-cell cloned, carried in culture for many generations to establish stability, and their karyotypes are determined by standard cytogenetic procedures and by hybridization to human DNA probes of known location. The human genome mapping project could begin by collecting the best-characterized cell lines for the chromosomes that are available. Some of the human chromosomes are not yet isolated in rodent lines due to the lack of selectable markers; brute force screening and new methods will likely make these available in the near future. If not, an early step in the chromosome mapping project will have be to make cell hybrids carrying these chromosomes. However, one point to consider is whether a single individual's genome should be mapped or whether the

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

map should consist of combined data from many different individuals. If the former course is chosen, the genome mapping project will require that new somatic cell hybrids carrying single human chromosomes be made from lymphoblasts isolated from the individual whose genome is selected for the mapping project.

The advantages of using somatic cell hybrids to obtain each human chromosome are that the amount of DNA that can be obtained is essentially limitless and that the chromosomes can be guaranteed pure of any other human material. In addition, many cell lines are already available and would be relatively easy to generate again if necessary. One disadvantage with this approach is that it does not isolate each human chromosome from the hamster or mouse chromosomes present in the hybrid cell. Thus, in the subsequent subdividing that occurs in Step 2 of the mapping experiment, a large amount of rodent DNA will be present. This is not a major problem since the human fragments can be distinguished from rodent by hybridization to human repeat sequences, for instance "Alu" (Jelinek et al., 1980) or "LINE" (Shafit-Zagardo et al., 1982) sequences. Alternatively, total human DNA can be used as probe to identify human sequences in the clones (Murphy et al., 1985). Another potential problem with using somatic cell hybrids as a source of human chromosomes is that occasionally chromosomal rearrangements occur when the hybrids are formed; the occurance of a rearrangement in a chromosome for which an enormous amount of effort will be expended in mapping experiments will seriously compromise the validity of the final map. One way to control for these rare problems is to obtain each chromosome from more than one cell hybrid.

The second method for obtaining purified human chromosomes is a technique that allows some of the chromosomes to be physically separated from the others. This method, known as chromosome sorting, or flow cytometry, uses fluorescent markers that bind to the chromosomes (Gray et al., 1986; Pinkel et al., 1986; Deaven et al., 1986; Lebo et al., 1986, and references therein). The fluorescently labeled chromosomes are then

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

dropped one by one in a small stream of liquid through a chamber; lasers direct a beam through the chamber that deflect the chromosomes to a degree dependent on the amount of fluorescent dye attached to them. Because human chromosomes differ in the degree to which they bind the fluorescent markers, it is possible to use this approach to physically separate many of them from all the others. Unfortunately, some chromosomes, particularly the larger ones, cannot be completely sorted; in some cases it is possible to separate some chromosomes into specific pairs. Even though complete separation of all the chromosomes would be ideal, partial purification is better than none, so some of the purified pairs will be useful. A second problem with this method of obtaining purified chromosomes is one of quantities; the amount of material obtainable by sorting is much lower than that obtained from somatic cell hybrids. Nevertheless, enough DNA can be obtained with the method to make reasonable phage or cosmid libraries, so this is probably not a big issue for the genome mapping project (see the above references).

Because of the shortcomings mentioned for each of the two procedures, it is likely that innovative combinations of both methods will be used in generating the isolated chromosomes for Step 1 of the mapping project. For instance, it may be desirable to use flow cytometry to separate the single human chromosomes present in somatic cell hybrids from the rodent chromosomes. The purification need not be complete, since even partial separation will decrease the number of non-human DNA fragments present in the next step (as discussed in the next section, partial purification will cut down the labor needed to screen through the subsequent clones for human as opposed to rodent inserts). This approach may be particularly important for separating human chromosomes that are not easily sorted when the others are present. In any case, both cell hybrids and sorting will likely be necessary to obtain each of the chromosomes in relatively pure form for the next step.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Step 2: Separation of each chromosome into fragments one to several million bp in length.

At this point, it would be possible to construct cosmid libraries from each of the purified human chromosomes, and then determine the order of each of the cloned inserts by identifying overlapping restriction fragments. However, as mentioned above, this approach would be complicated by the enormous size of the chromosomes, even the smallest ones. The average size of a chromosome is 10^8 bp; a near-complete cosmid library of an average chromosome would contain over 10,000 cosmids [12,500 cosmids, actually, to achieve 5-fold redundancy]. Ordering this number of cosmids by overlapping restriction fragments really is not feasible. It would be much more practical to attempt the cosmid ordering on libraries containing fewer members. Thus, if each chromosome could be divided into about 50 to 200 fragments (average size 0.5 to 5 million bp), and then cosmid libraries are made from these smaller fragments, the problem of ordering cosmid inserts is greatly diminished; determining the order of 50 to 200 inserts in each library by overlapping restriction fragments is a task that can be performed in a straightforward manner. Note, however, that a tradeoff in labor is made with this approach; a large number of cosmid libraries will have to be made. Although construction of libraries and screening for human inserts requires a good deal of labor, subdivision of each chromosome at this step will greatly ease the task of ordering the cosmid inserts, and if done properly, will actually allow the order of restriction sites within each insert to be determined in the same step (see Step 4).

How can chromosomes be divided into the fragments needed for this step? Currently, two methods are available; both of these approaches are new technologies. The first, which can be used to physically isolate DNA fragments in this size range, involves an electrophoretic system called pulsed-field gel electrophoresis (PFG). A

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

second, recently developed procedure allows one to clone foreign DNA fragments that are hundreds or thousands of kb in length into yeast. The end result is a library of yeast strains, each of which carries a single foreign DNA fragment on a separate "artificial" chromosome. The library, if large enough, will carry a set of overlapping DNA fragments that comprise the entire human chromosome from which it was made.

The pulsed field gel electrophoresis system was first described by D. Schwartz and C. Cantor (Schwartz and Cantor, 1984) and subsequently modified and improved by M. Olson and colleagues (Carle and Olson, 1985; Carle et al., 1986) and R. Davis and colleagues (Chu et al., 1986). These gel systems allow separation of DNA molecules in the size range 20,000 to 10,000,000 bp and in some cases somewhat larger. The description of this method understandably caused much excitement in the research community; before its invention, the upper limit of resolution (by standard agarose gel electrophoresis) was about 10,000 bp. This new method is so powerful it can be used to separate all 14 of the S. cerevisiae chromosomes from each other. Essentially, carefully prepared (i.e. non-sheared) DNA fragments are subjected to electrophoresis in an agarose gel that has been altered so that the electric field driving the electrophoresis of the molecules is switched back-and-forth during the run. For some poorly understood reason, this alteration in the electric field allows very large DNA molecules to migrate into the gel according to their molecular weights rather than displaying the normal limited mobility of all molecules above 10,000 bp. The result is remarkable resolution of DNA fragments over a large size range. Several variations of the method exist; these largely are modifications in the switching times or the configuration of the electric field. The PFG system, although new, is "tried and true"; a number of important research papers, particularly in human genetics, have been published describing its use to answer physical mapping problems similar to the ones that will be encountered in the human genome project (Monaco et al., 1986; Darling et al., 1986; Burmeister et al., 1986; Kenrick, et al., 1987).

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

The PFG system is useful in physical mapping of large regions of chromosomes, in the range of 0.5 to perhaps tens of millions of bp. It can be used analytically to determine the physical relationships of several large DNA fragments, which are generated by cleavage of intact chromosomes with rare-cutting restriction enzymes such as NotI and SfiI. The methods for mapping NotI or SfiI sites in a chromosomal region are very similar to those used to determine a restriction map on a small scale, such as that of a virus genome, except that the sizes of the DNA fragments are 10 to 100 times larger. In addition, the gels can be used preparatively to purify large DNA fragments (Michiels et al., 1986). DNA fragments of 100,000 to several million bp can be generated from genomic DNA samples by digestion with NotI or SfiI, and the fragments can be separated by PFG electrophoresis, followed by elution of specific fragments from the gel. Eluted fragments can then be digested with a more frequent-cutting enzyme into fragments of about 40,000 bp for cloning into cosmid vectors as described below.

Thus, in principle, one may be able to separate each isolated chromosome into about 100 to 200 fragments with the PFG system, then selectively isolate DNA from each of these fragments for cosmid cloning fine-structure mapping. However, a limitation with this approach is the resolution of the PFG system; even though large DNA fragments can be separated on the system, it is unlikely that more than 15 to 20 bands will actually be resolved on a pulsed field gel. Therefore, instead of separating each chromosome into 100 to 200 fragments on the gel, most bands will actually represent an average of five to ten different DNA fragments which are very likely not linked in the intact chromosome. Purification of a band, followed by cosmid cloning and physical mapping, will in fact not be very helpful then, since the DNA from each band represents a fairly large portion of the chromosome. For this reason, it seems unlikely that PFG will be used to purify fragments from whole chromosomes in the genome mapping project. Nevertheless, it is likely that the gel system will be useful in combination with the second approach described below (the yeast artificial chromosome system) and for

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

confirming the long-range map of each chromosome that will be developed in Step 5.

A very recent development, again provided by Maynard Olson and his colleagues, has caused much excitement among those scientists engaged in discussing the human genome mapping project and in the field of genetics in general (Burke et al., 1987). This method, which has only just been submitted for publication, provides a way to generate artificial chromosomes in yeast that carry sections of exogenous DNA ranging in size from about 100,000 bp to over a million bp. This development is extremely important for the field of human genetics; until now, the largest size fragments that could be cloned molecularly have been on the order of 40,000 bp. Besides greatly aiding the human genome mapping project, this method will make significant contributions for researchers working on practically any aspect of chromosome structure and function from any organism. Although the largest artificial chromosome made by Olson's group during the development of the procedure is slightly over one million bp, it is likely that modifications of the procedure will allow fragments up to ten million bp to be cloned as artificial yeast chromosomes.

The yeast cloning system is simple in concept and experimentally. Genomic DNA from the organism in question is isolated by gentle methods that do not shear the DNA, and this DNA is treated with a restriction enzyme such as EcoRI under conditions that result in extremely partial digests (such as one EcoRI cleavage per million bp; EcoRI sites occur on average every 6,000 bp in most genomes). These very large EcoRI partial fragments are mixed with a small (5,000 bp) plasmid vector containing a yeast centromere sequence flanked by two yeast telomere sequences (Figure 2). The vector has also been digested with EcoRI, which cleaves between one of the telomeres and the centromere. The EcoRI ends of the genomic DNA are annealed to those in the vector and the mixture is treated with DNA ligase to connect the insert into the vector to form a large covalently closed circle. The ligated DNA is then introduced by transfection into

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

yeast; individual yeast cells picking up and stably replicating one of the artificial chromosomes are selected by a marker present on the cloning vector. This approach was used to generate hundreds of yeast strains carrying different fragments of human DNA in the size range 100,000 bp to 1,100,000 bp. The efficiency of cloning artificial chromosomes by this method is on the order of several hundred transfectants per microgram of vector; these efficiencies are clearly good enough to construct yeast libraries carrying all of the human fragments present in a purified chromosome or even in a somatic cell hybrid carrying a single human chromosome.

There are technical difficulties in generating inserts of several million bp, so this has to be a consideration for a massive genome mapping project. In particular, special precautions must be taken to prevent shearing of the genomic DNA during its preparation. As mentioned above, the technology is already good enough for inserts in the one million bp range, which means that an average human chromosome would be represented by a minimum of 100 yeast strains, a high but feasible number for the project. However, I believe that with some modifications, inserts larger than one million bp can be made; for instance, it may be possible to prepare the inserts in agarose blocks (the current method used to prepare very large [up to 14 million bp] intact genomic DNA fragments for PFG). If so, this will consequently decrease the number of yeast strains that will have to be constructed to achieve complete libraries.

The yeast strains constructed by this method can be grown on nitrocellulose filters and then their DNA probed in situ for the presence of human sequences (Fangman and Dujon, 1984). Those strains carrying human artificial chromosomes (perhaps even selected for those with the largest human inserts based on strength of hybridization to human repeat sequences; see Porteous, 1985 and Murphy et al., 1985) can be identified. The size of the human artificial chromosomes in these strains can be determined by PFG. A number of the yeast strains predicted to contain all the human sequences

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

present in the starting intact human chromosome are selected for further analysis. Genomic DNA is isolated from these strains, and this is then cloned into a cosmid vector as described in Step 3.

As discussed above, it seems likely that both PFG and the yeast cloning system will need to be used for this step of the project. The PFG system alone will not be effective for purification of 0.5-5 million bp human fragments due to the problem of overlapping fragments in a single band, as mentioned above. However, in some cases it may be necessary to use PFG to obtain single bands representing several different fragments. For instance, it is possible that some human DNA fragments in this size range cannot be cloned as artificial yeast chromosomes [note that there is no evidence that such is the case]; it this happens, the "missing" fragments can be obtained from PFG bands.

Step 3: <u>The division and isolation of each chromosomal subfragment into 40,000 bp overlapping fragments: Construction of cosmid libraries.</u>

This step is the most straightforward one in the entire procedure. Genomic DNA isolated from each of the yeast strains above is digested with a frequent-cutting restriction enzyme, one that cuts on average every 500 bp or so; however, the digestion is not done to completion, but instead, a "partial" digest is performed by lowering the concentration of the restriction enzyme. The partial conditions are adjusted so that DNA fragments of average size of 40,000 bp are generated. Because the cleavage sites are chosen by the enzyme in a relatively random fashion, the resulting DNA molecules will consist of a population of overlapping fragments representing all of the DNA in the yeast genome, including the artificial chromosome carrying the human fragment. These fragments are cloned into a cosmid vector to obtain a library, which is a collection of thousands of colonies whose 40,000 bp inserts collectively contain the entire genome several times over. Duplicates of each bacterial plate are made onto nitrocellulose

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

filters, and one of the duplicates is lysed in situ, and hybridized to human "Alu" repeat sequences to identify those colonies carrying a cosmid with a human insert (Jelinek et al., 1980). Because Alu sequences appear on average every 5,000 bp in the human genome, the great majority of human fragments 40,000 bp in length will contain several of them. A small fraction of human cosmid inserts do not carry an Alu repeat sequence; these regions of "Alu-minus" DNA are probably not more than 50,000 bp in length and are very rare (one example of a 40,000 bp Alu-minus stretch is the G8 probe linked to the Huntington's Disease locus on chromosome 4; Gusella et al., 1983). Thus, some of the yeast artificial chromosomes from which the cosmid libraries are made may have a stretch of DNA, up to 50,000 bp perhaps, lacking Alu repeats; such a case will result in a gap in the restriction map, determined by the methods of Step 4, of the artificial chromosome. In such a case, the gap can easily be filled by isolating cosmids that are homologous to DNA sequences at both ends of the gap. This experimental approach is similar to that used in "cosmid walking" (Evans and Wahl, 1987).

Although the construction of cosmid libraries and screening for inserts by in situ hybridization to colonies is routine, a few words about numbers is in order here. If this approach is used to divide each chromosome into 50 to 200 subchromosomal fragments in the yeast libraries described above, then between 1000 and 4000 cosmid libraries will have to be constructed and screened to represent the entire human genome. Clearly this is an enormous amount of work, but it is important to remember that this division and subdivision will greatly decrease the efforts needed to order the cosmids and determine their restriction maps. In fact, it is unlikely that the mapping can be done without some degree of subdivision at the beginning. Therefore, the investment of a relatively large amount of effort to prepare the materials for the fine-structure restriction mapping will pay off in the next difficult step. One technical point should be made: The screening for cosmids with human inserts very simple in this scheme since the yeast genomes from which the libraries are made are of low complexity. Based on a yeast genome size of 1.5

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

x 10^7 bp (Lauer et al., 1977) and the size of the human fragments present in the artificial chromosomes, human sequences will be present in about 3 percent to 30 percent of the cosmids; therefore, no special efforts will be needed to construct and screen the cosmid libraries to ensure that they are complete.

Step 4: Determining the order and restriction maps of the 40,000 bp DNA fragments.

The next step is a difficult one: the difficulty comes from the large amount of labor that will be required to order and map the thousands of cosmids representing the whole human genome. The purpose of this step is to align the cosmid inserts into the order that they appear in the genome, and then to determine the patterns of cleavage sites for one or two restriction enzymes within each of the inserts. This process is essentially that of "restriction mapping", which has been an important part of molecular biology since the discovery of restriction enzymes in the 1970s. On a small scale, restriction mapping is easy. For instance, in many cases, it is desirable to construct a restriction map of a small cloned region of DNA, say a 10,000 bp insert in a bacterial plasmid. Generally, the positions of cleavage sites within the insert for several restriction enzymes is determined by digestion of aliquots of the DNA with each enzyme separately, and with pairs of each enzyme in "double digests." The restricted DNA is separated by size on an agarose gel and visualized by ethidium bromide staining. By comparing the sizes of all the fragments in each digest, it is possible to deduce, usually without ambiguity, the relative positions of each of the cleavage sites within the insert. In ambiguous cases, the information gained from the ethidium-stained pattern can be supplemented by blotting the gel and using subfragments of the insert as hybridization probes to the blot. Alternatively, it is often possible to overcome ambiguities by purifying subfragments of the insert and simply determining the restriction map of each subfragment, followed by compilation of the results into a complete restriction map of

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

the whole insert.

The human genome project will require restriction mapping on a scale never before attempted. It will be a large task to order the cosmids and determine their restriction maps from even a single yeast artificial chromosome; this task will be amplified over 1,000 fold in order to analyze all the artificial yeast chromosomes representing the entire human genome. Nevertheless, the manipulations for this step are straightforward. Two groups have devised methods for achieving this objective; Maynard Olson and his colleagues for the yeast genome (Olson et al., 1986) and a group at the MRC, Cambridge for the nematode genome (Coulson et al.; 1986). Their efforts have yielded maps for 60 to 80 percent of the two genomes in just a few years; many of their ideas and improvements will be important for the human genome project.

The approach designed for the yeast and nematode genomes allows the order of the cosmid inserts and the restriction enzyme cleavage sites within the inserts to be determined simulataneously. A similar analysis will be used for the human genome mapping study; however, the ordering is done one at a time on the cosmid libraries made from each of the subgenomic fragments (i.e. artificial yeast chromosome carrying a human DNA insert) in the yeast strains. As discussed above, hundreds of cosmids with human inserts are isolated from each yeast strain carrying a human artificial chromosome; these cosmids represent all the human sequences in the yeast strain with several-fold redundancy [some numbers: average size yeast artificial chromosome carrying human DNA = 2×10^6 bp; average cosmid insert = 4×10^4 bp; so 50 cosmids would represent the entire subfragment; for 5-fold redundancy (to achieve a high degree of overlap), we would need an average of 250 cosmids carrying human inserts from the library].

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

The experimental manipulations for ordering and determining the restriction maps for 250 cosmids are shown in Figure 3; note that the approach is essentially that developed by Olson's group for mapping the yeast genome (Olson et al., 1986) and by the MRC group for the nematode genome (Coulson et al., 1986). A modification of the approach has recently been used successfully to find a candidate gene for the cystic fibrosis locus (Estivill et al., 1987). The DNA inserts of each cosmid clone are digested with a restriction enzyme (or possibly a pair of enzymes) that cuts the inserts into many fragments. [Note: in Olson's study, the restriction enzymes (a mix of EcoRI and HindIII) used cut the phage inserts into fragments with an average size of about 4,000 bp. In the Coulson et al. approach, especially shown by Estivill et al., a much larger number of fragments can be generated by digesting the cosmid DNA with a frequent-cutter, such as HinfI (average size about 500 bp), end-labeling the fragments, and analyzing the 50 to 100 fragments on a denaturing acrylamide gel. This approach allows a much finer resolution map to be made than with the EcoRI/HindIII cutting. Clearly this second approach would be preferable for the human genome mapping project, since a much larger amount of information can be gathered for approximately the same amount of effort.] Then the fragments are separated according to size by electrophoresis in an agarose or acrylamide gel. The DNA fragments are visualized by ethidium bromide staining, and a high quality photograph of the gel is made. Alternatively, the DNA has been end-labelled and the restriction fragments are visualized by autoradiography. Each clone will have a unique or almost unique set of restriction fragments; thus the gel pattern is something like a fingerprint of each insert. A comparison is made of the patterns of restriction fragments for each of the clones to search for inserts that contain some overlapping fragments. Clones that are found to share some bands are very likely to be contiguous in the subchromosomal fragment, and consequently in the human genome. The collection of the data, which consists of restriction bands on the photographs or autoradiograms of the gels, can be done by digital scanning of the

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

photographs, followed by computer-assisted analysis of the data. This technology for data analysis is already well-established from Olson's work with the yeast genome (Olson et al., 1986) and from protein studies as well (Skolnick et al., 1982; Skolnick, 1982; Rosenblum et al., 1983).

Since the cosmid library will represent the human sequences present in the yeast strain by several-fold redundancy, most human sequences will be present several times in different overlapping cosmid inserts. If enough overlaps are available, it is possible to use this data to map not only "contigs" (cosmid inserts that appear next to each other, and slightly overlapping, in the human genome), but also to map the order of the restriction enzyme cleavage sites within the cosmid inserts. This method of course is greatly simplified by the computer analysis. In addition, the analysis will actually be somewhat easier for each of the cosmid libraries analyzed in the human genome project than it was for the yeast and nematode mapping projects, since each human library will represent only a few million bp of human DNA at most. By comparison, the haploid yeast genome contains 1.5×10^7 bp (Lauer et al., 1977) and the nematode genome contains 8×10^7 bp (Sulston and Brenner, 1974).

At the end of this step in the procedure, we will have complete restriction maps (that is, a linear order and position of two restriction enzyme cleavage sites) of each of the artificial chromosomes carrying human DNA from the yeast strains. With this information, the map is almost complete; what remains is to link up each of the human subchromosomal fragments in the order that they appear in the intact chromosome (Step 5).

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Step 5. Linking up the restriction maps of each cosmid: Completion of the map.

This last step would in fact not be a separate procedure to be done at the end of the project, but instead would be an ongoing process throughout the progression of the project as the restriction data is gathered. As in all saturation mapping projects, during the early course of the mapping, small "islands" of linked-up sequences will be established. These islands will become larger and larger as more data is gathered, until finally they are all connected into a continuous map of an entire chromosome. Ultimately, the complete map will contain continuous maps of each of the twenty-two autosomes and the two sex chromosomes.

By using the same principles discussed above to determine the order of the cosmids within a single 0.5-5 million bp artificial chromosome, the restriction site and cosmid-ordering data gathered in Step 4 can be used to determine the areas of overlap between the artificial chromosomes. Again, the overlaps can be established by computer comparison of the restriction maps of all the cosmids analyzed from each artificial chromosome. In a sense, this step will be a continuation of Step 4, since the data (i.e., cosmid restriction maps) will have already been input into the computer. In fact, as soon as cosmid-ordering data from several artificial chromosomes is obtained, this step could be started, resulting in the linking up of some 0.5-5 million bp segments early on. This type of analysis will ultimately allow the human DNA within each of the artificial chromosomes to be aligned in the order that it appears in each of the human chromosomes; the end result will be the final physical map of the human genome.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

DISCUSSION

The experimental approach outlined above is not written in stone; in fact, it is likely that the actual project will use variations of the methodology described here. However, with today's technology, the mapping project will require the basic strategy outlined here: Cloning of the entire genome in some vector and determination of the restriction maps of each of the clones, followed by linking up all the restriction maps into a single map of the whole genome. It may be possible that either more or less subdivision of the genome prior to mapping will make the mapping more feasible, although most reports and informal discussions amongst researchers favor a subdivision to fragments in the size range of a few hundred kb to several million bp. Similarly, other cloning vectors that offer improvements over cosmids may be derived in the near future, so it is possible that they would be used in the human genome mapping project. These and other modifications are details that may not be decided until the project is underway; we should probably proceed with something like the approach described here and modify when necessary.

No mention of automation in the mapping strategy has been made. Clearly, because of the enormous amount of labor that will be involved in the many repetitive steps in the project, any degree of automation will help lower the overall costs. Several robotic devices have been designed and used successfully by recombinant DNA companies (Genetics Institute, Inc. and Cetus, for example) for repetitive DNA cloning steps, such as preparation and restriction digestion of plasmid and cosmid DNA. Similarly, there is discussion about automating the pouring, loading, and running of agarose gels. It is easy to see how many of the steps in the above strategy could be adapted to automation (other cloning procedures, probe synthesis, and Southern blotting of the cosmid colonies, for example), and these approaches should be explored by the groups planning the project.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Because technical advances are appearing so rapidly in the fields of biology, robotics, and computers, it is difficult to imagine what the near future will bring. It is quite possible that some as yet-developed technology will make many of the current physical mapping procedures obsolete very quickly. For instance, automated DNA sequencing, which is now feasible at a rudimentary level (Lewin, 1986a), may be advanced to the point that it would be reasonable to map the human genome by random sequencing of clones representing the entire genome. In fact, several fantastic predictions have been made for the number of bp that can be sequenced by the machines of the near future (see Lewin 1986, for example); if such predictions become reality, we should reevaluate the above approach and map by random sequencing (such an approach would obviate the need to perform all the preliminary subdivisions discussed in this report). However, because no rapid automated sequencers have been demonstrated yet, we should begin the project with a more conventional mapping approach such as that described here. A major function of the groups coordinating the mapping project should be to support the development of these new technologies so that they can be tapped immediately when they are ready. We have reason to be optimistic about new developments; the explosive growth already present in biotechnology will very likely be amplified when a major concerted effort like the human genome mapping project is undertaken.

Many other questions need to be addressed before a major effort is undertaken to map the human genome. One important consideration is whether to begin by concentrating the first mapping efforts on regions of the chromosomes containing known genes, or to begin physical mapping of the entire genome at the start of the project. Secondly, regardless of the order mapping particular regions of the genome, consideration of how to incorporate the locations of known and unknown genes (defined both by cloned genes and by genetics) into the physical map. This second point is particularly important, since one of the major reasons for mapping and sequencing the

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

human genome is to use the information and DNA clones to study and develop treatments for the 3,500 genetic diseases affecting humans. It would be silly to charge through the physical mapping project without using our knowledge of the several hundred genes that have been cloned or identified genetically; by beginning the physical mapping around these genes, immediate benefits will be seen.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

REFERENCES

Bernard, H.U., and Helinski, D.R. (1980) Bacterial plasmid cloning vehicles. In Genetic Engineering (ed. J.K. Setlow and A. Hollaender; Plenum Press, N.Y.) volume 2, pp. 133-xx.

Bolivar, F., and Backman, K. (1979) Plasmids of E. coli as cloning vectors, Methods in Enzymology 68, 245-xx.

Boone, C., Chen, T.-R., and Ruddle, F.H. (1972) Assignment of three human genes to chromosomes (LDH-A to 11, TK to 17, and IDH to 20) and evidence for translocation between human and mouse chromosomes in somatic cell hybrids. Proc. Natl. Acad. Sci. USA 69:510-514.

Botstein, D., White, R.L., Skolnick, M., and Davis, R. W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32, 314-331.

Burke, D.T., Carle, G.F., and Olson, M.V. (1987) Cloning of large segments of exogenous DNA into yeast using artificial-chromosome vectors. Science, In press (due to appear in late May).

Burmeister, M., and Lehrach, H. (1986) Long-range restriction map around the Duchenne muscular dystrophy gene. Nature 324, 582-585.

Carle, G.F., and Olson, M.V. (1985) An electrophoretic karyotyoe for yeast. Proc. Natl. Acad. Sci. USA 82, 3756-3760.

Carle, G.F., Frank, F., and Olson, M.V. (1986) Electrophoretic separations of large DNA molecules by periodic inversion of the electric field. Science 232, 65-68.

Chehab, F.F., Honig, G.R., and Kan, Y.W. (1986) Spontaneous mutation in beta-thalassemia producing the same nucleotide substitution as that in a common heriditary form. The Lancet, January 4, 1986, pp. 3-5.

Clarke, L., and Carbon, J. (1976) A colony bank containing synthetic ColE1 hybrid plasmids representative of the entire E. coli genome. Cell 9, 91-99.

Coulson, A., Sulston, J., Brenner, S., and Karn, J. (1986) Toward a physical map of the genome of the nematode Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA 83, 7821-7825.

Deaven, L.L., et al. (1986) Construction of human chromosome-specific DNA libraries from flow-sorted chromosomes. CSHSQB 51, 159-167.

Estivill, X., et al. (1987) A candidate for the cystic fibrosis locus isolated by selection for methylation-free islands. Nature 326, 840-845.

Evans, G.A., and Wahl, G.M. (1987) Cosmid vectors for genomic walking and restriction mapping. Methods in Enzymology, vol. 152: A Guide to Molecular Cloning. In press.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Fangman, W.L., and Dujon, B. (1984) Yeast mitochondrial genomes consisting of only A.T pairs replicate and exhibit suppressiveness. Proc. Natl. Acad. Sci. USA 81, 7156-7160.

Fischer, S.G., and Lerman, L.S. (1983) DNA fragments differing by single base pair substitutions are separated in denaturing gradient gels: Correspondence with melting theory. Proc. Natl. Acad. Sci. USA 80, 1579-1583.

Gall, J.G. (1986) letter to the editor in Science 233, 1367-1368.

Glass, R.E. (1982) Gene Function: E. coli and its Heritable Elements (University of California Press, Berkeley, CA).

Gray, J.W., et al. (1986) Flow karyotyping and sorting of human chromosomes. CSHSQB 51, 141-149.

Gusella, J.F., Wexler, N.S., et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306, 234-238.

Herskowitz, I., (1973) Control of gene expression in bacteriophage lambda. Annu. Rev. Genet. 7, 289-xx.

Hendrix, R.W., Roberts, J.W., Stahl, F.W., and Weisberg, R.A. (1982) Lambda II. (Cold Spring Harbor Press, Cold Spring Harbor, N.Y.)

Hohn, B. and, Collins, J. (1980) A small cosmid for efficient cloning of large DNA fragments. Gene 11, 291-xx.

Ish-Horowicz, D., and Burke, J.F. (1981) Rapid and efficient cosmid cloning. Nucleic Acids Res. 9, 2989-2998.

Jelinek, W.R., et al. (1980) Ubiquitous, interspersed repeated sequences in mammalian genomes. Proc. Natl. Acad. Sci. USA 77, 1398-1402.

Kenrick, S., Patterson, M., Speer, A., Fischbeck, K., and Davies, K. (1987) Molecular analysis of the Duchenne muscular dystrophy region using pulsed field gel electrophoresis. Cell 48, 351-357.

Lauer, G.D., Roberts, T.M., and Klotz, L.C. (1977) J. Mol. Biol. 114, 507-526.

Lebo, R.V., Anderson, L.A., Lau, Y-F.C., Flandermeyer, R., and Kan, Y.W. (1986) Flow-sorting analysis of normal and abnormal human genomes. CSHSQB 51, 169-176.

Lewin, R., (1986a) Proposal to sequence the human genome stirs debate. Science 232, 1598-1600.

Lewin, R., (1986b) Shifting sentiments over sequencing the human genome. Science 233, 620-621.

Lewin, R., (1987) National academy looks at human genome project, sees progress. Science 235, 747-748.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) in Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Press, Cold Spring Harbor, N.Y.) pp. 270-274.

McKusick, V.A. (1986) The gene map of Homo sapiens: Status and Prospectus. CSHSQB 51, 15-27.

Michiels, F., Burmeister, M., and Lehrach, H. (1987) Derivation of clones close to the cystic fibrosis marker MET by preparative field inversion gel electrophoresis.

Monaco, A.P., Neve, R.L.m Colletti-Feener, C., Bertelson, C.J., Kurnit, D.M., and Kunkel, L.M. (1986) Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. Nature 323, 646-650.

Murphy, P.D., Miller, C.L., and Ruddle, F.H. (1985) Quantitation of the transgenome size in chromosome-mediated gene transfer lines. Cytogenet. Cell Genet. 39, 125-133.

Myers, R.M., Lumelsky, N., Lerman, L.S., and Maniatis, T. (1985a) Detection of single base substitutions in total genomic DNA. Nature 313, 495-498.

Myers, R.M., Larin, Z., and Maniatis, T. (1985b) Detection of single base substitutions by ribonuclease cleavage at mismatches in RNA:DNA hybrids. Science 230, 1242-1246.

Myers, R.M., and Maniatis, T. (1986) Recent advances in the development of methods for detecting single-base substitutions associated with human genetic diseases. CSHSQB 51, 275-284.

Noll, W. and, Collins, M. (1987) Proc. Natl. Acad. Sci. USA. In press.

Olson, M.V., Dutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., Maccollin, C., Scheinman, R., and Frank, T. (1986) Random-clone strategy for genomic restriction mapping in yeast. Proc. Natl. Acad. Sci. USA 83, 7826-7830.

Pinkel, D., et al. (1986) Cytogenetic analysis by in situ hybridization with fluorescently labeled nucleic acid probes. CSHSQB 51, 151-157.

Porteous, D.J., (1985) Rapid and quantitative detection of unique sequence donor DNA in extracts of cultured mammalian cells: An aid to chromosome mapping. Somatic Cell and Molecular Genetics 11, 445-454.

Rosenblum, B.B., Neel, J.V., and Hanash, S.M. (1983) Two-dimensional electrophoresis of plasma polypeptides reveals "high" heterozygosity indices. Proc. Natl. Acad. Sci. USA 80, 5002-5006.

Royer-Pokora, B., Kunkel, L.M., Monaco, A.P., Goff, S.C., Newburger, P.E., Baehner, R.L.., Cole, F.S., Curnutte, J.T., and Orkin, S.H. (1986) Cloning the gene for an inherited human disorder - chronic granulomatous disease - on the basis of its chromosomal location. Nature 322, 32-38.

Schwartz, D.C., and Cantor, C.R. (1984) Separation of yeast chromosome-sized DNAs by pulsed field gel electrophoresis. Cell 37, 67-75.

Shafit-Zagardo, B., Maio, J.J., and Brown, F.L. (1982) KpnI families of long, interspersed repetitive DNAs in human and other primated genomes. Nucleic Acids Res. 10, 3175-3193.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Shay, J.W., (1982) Techniques in Somatic Cell Genetics (Plenum Press, N.Y.).

Sinsheimer, R.L. (1986) letter to the editor in Science 233, 1246.

Skolnick, M.M., Sternberg, S.R., and Neel, J.V. (1982) Computer programs for adapting two-dimensional gels to the study of mutation. Clinical Chemistry 28, 969-978.

Skolnick, M.M., (1982) An approach to completely automatic comparison of two-dimensional electrophoresis gels. Clinical Chemistry 28, 979-986.

Smith, D.A., (1986) letter to the editor in Science 233, 1246.

Solomon, E., and Bodmer, W.F. (1979) Evolution of sickle variant gene. The Lancet i, 923.

Sulston, J.E., and Brenner, S. (1974) Title. Genetics 77, 95-104.

Vogel, F., and Motulsky, A.G. (1986) Human Genetics: Problems and Approaches (Springer-Verlag, N.Y.) pp. 369-370.

Williams, B.G., and Blattner, F.R. (1980) Bacteriophage lambda vectors for DNA cloning. In Genetic Engineering (ed. J.K Setlow and A. Hollaender) volume 2, pp. 201-xx.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Appendix 1: GLOSSARY OF TERMS

Bacteriophage (often shortened to "phage"): a bacterial virus, composed of a single species of DNA or RNA encapsidated in a protein shell. Some phage DNAs have been altered for use as molecular cloning vectors. Bacteriophage lambda is a particularly useful cloning vector; it can accomodate foreign DNA inserts up to 20,000 bp.

Clone: for the purposes of this report, "clone" refers to a molecular clone, which is a segment of DNA that has been isolated from the human genome and specifically amplified to many copies in a bacterial host. The amplification can be performed only if the human DNA fragment is first covalently attached to another, generally smaller DNA fragment called a vector. The term "cloning" is often given to the process of isolating genomic DNA, digesting it with restriction enzymes, covalently attaching it to a vector DNA sequence, introduction of this DNA into bacteria, and selection of individual bacterial strains, each carrying a single species of amplified DNA. Note that cloning of specific DNA molecules can also be done in DNA vectors that replicate in other (non-bacterial) organisms such as yeast and even mammalian cells. In fact, the procedure in this paper describing the isolation of artificial chromosomes carrying human DNA fragments in yeast is one way of cloning in yeast.

Cosmid: a cosmid is a type of bacterial DNA vector, usually about 5,000 bp in length, that can accommodate DNA inserts in the size range of 35,000 to 40,000 bp.

Genetic map: a genetic map is a depiction of the order along the lengths of chromosomes of specific genes. Genes are defined by specific phenotypes whose inheritances can be followed by Mendelian genetics. Their order is determined by the degree to which they recombine with one another during meiosis. A genetic map can give a rough idea of the physical distances between genes if the size of the genome is known; however, these measures of distance are subject to error since the rate of recombination is not linear along the length of chromosomes (i.e., there are "hot" and "cold" spots).

Genome: the sum total of the DNA present in a single cell of an organism. A distinction is usually made between the nuclear genome and mitchondrial genome (which is quite small) in eukaryotic cells. Note that the genome is not fixed throughout all cell types in complex organisms. For instance, DNA rearrangements are known to occur in the regions of mammalian genomes coding for immune functions; other rearrangements may occur in different tissues and chromosomal regions.

Karyotype: the entire chromosomal complement of a cell or species, usually visualized by spreading mitotic chromosomes on a microscope slide and examined by staining the chromosomes with specific dyes.

Physical map: a physical map is a linear depiction of the order and distance between specific markers (expressed as base pairs, restriction enzyme cleavage sites, genes, or segments of DNA of unknown function) along the length of a region of DNA. The resolution of most physical maps is at the level of restriction sites; the ultimate physical map of a region is its compltete nucleotide sequence. A physical map can be determined for small DNA segments, such as plasmids, or very large DNA fragments, such as chromosomes. The largest contiguous region of human DNA that has been physically mapped is about one million bp.

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Plasmid: a DNA vector, between 1,000 and 10,000 bp in length, that can carry foreign DNA inserts up to 12,000 bp. Plasmids grow to high copy number in bacteria by confering resistance to bacteriocidal drugs.

Vector: in this paper, the term "vector" means a specific DNA molecule, usually in the form of a small (<10 kb) circular plasmid or a linear bacteriophage genome, that can be used to clone a specific foreign DNA fragment called an "insert". Vectors carry in their DNA sequence information for propagation in a bacterial, yeast, or even higher eukaryotic cell; this information includes such as an origin of replication, a gene or two that allows selection for drug resistance, and a "polylinker" sequence which allows the cloning of the foreign insert into a nonessential part of the vector [Note: No higher eukaryotic vectors are discussed in this report].

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Appendix 2: Genetic Maps

A genetic map is a depiction of the linear order along the length of chromosomes of genes. As discussed in the text, the relative distances between genes can also be estimated by genetic mapping techniques; however, genetic distances often are not strictly correlated with physical distances since recombination frequencies are not linear along the lengths of chromosomes. A partial genetic map of the human genome has been constructed over the years by a number of techniques; we now know the rough chromosomal locations of 900 genes (McKusick, 1986). We still have a long way to go, since there are about 50,000 genes in our genomes. The physical map of the human genome will be much more valuable when genes can be assigned to the ordered DNA clones that are mapped physically; if some attention is paid to known areas containing genes at the onset of the physical mapping project, immediate medical benefits (in the form of diagnosis of genetic diseases) will ensue.

Genetic maps can be constructed on the basis of phenotypes; if two traits can be shown to segregate together during meiosis to a greater degree than they would by chance, then the traits are said to be genetically linked. Genetic linkage can be established with more and more certainty as the sample size becomes large enough. The order and relative genetic distances between traits can be determined by comparing cross-over frequencies between each trait; as more and more traits are added to the map, it becomes possible to place a new trait on the map due to its segregation behavior compared to two or three markers on the map, and then infer its position relative to all the other markers.

Many genes in humans have been genetically mapped by following their inheritance with already mapped traits that are observed biochemically, such as red blood cell surface antigens. However, such tests have limited value because there is relatively little variation in these antigens in the population; linkage of a new phenotype to an mapped marker can only be done if the mapped marker shows heterozygosity in the families being tested (this problem is discussed below). The discovery that human beings have a high degree of difference in their nucleotide sequences (estimated to be between 0.1% and 1%) lead to the notion that these DNA polymorphisms could be used as linkage markers instead of relying on easily observable phenotypes (Botstein et al., 1980; Solomon and Bodmer, 1979). This prediction has been born out; in just seven years, many genes in humans and other organisms have been mapped genetically due to their segregation with DNA fragments that are polymorphic within the population. Note that most of these polymorphic DNA markers are not genes, but simply randomly isolated fragments of DNA that show a high enough degree of sequence difference in the population that most families are heterozygous for the sequence difference.

Most polymorphisms in DNA fragments have been detected by observing the loss or gain of a restriction enzyme cleavage site within the genomic DNA of some of the individuals being tested. The loss or gain of the site is usually due to a single base substitution at one of the nucleotide positions within the recognition site for the enzyme; however, in some cases, polymorphisms are caused by insertions or deletions within a region. Alteration in the cleavage sites are detected by digesting genomic DNA samples from several individuals, electrophoresing the samples in an agarose gel, transferring the DNA to filter paper or nylon by the Southern blotting procedure, and then detecting the specific DNA sequence on the blot by hybridization to the cloned DNA marker that has

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

been radioactively labeled. Polymorphisms are indicated by a difference in the size of the probed DNA sample in some of the individuals; for this reason, the procedure is known as RFLP mapping (Restriction Fragment Length Polymorphism). The method is used in linkage studies by finding a probe that is reasonably polymorphic within the population [usually this means that at least 25% of the population carries one form, say a HindIII fragment hybridizing to the probe that is 5,000 bp, and 75% of the population carries a different form, say, the HindIII fragment now cut into two fragments, 2,000 bp and 3,000 bp. Probes that show less variation than 25/75 are still often useful; however, the less polymorphic a probe is, the larger the number of families that have to be tested to establish linkage], and then following the segregation of the polymorphism with the disease trait within a given family. If the parents have one difference in the probe (that is, one of the affected or carrier parent's alleles is different from both alleles in the other parent), then it may be possible to show that this difference is inherited with the disease in one or more of the offspring. Note that a particular form of the polymorphism (say, the presence of a HindIII site) that segregates with the disease in one family may not segregate with the disease in another family (but instead, in the second family, the absence of the particular HindIII site segregates with the disease). This is due to the fact that the polymorphic marker is not actually at the disease locus, but some distance away, and that the disease may have arisen many generations before. Thus, enough cross-overs between the disease and the marker region have occurred that both alleles of the marker have a significant likelihood of being on the same chromosome as the mutant gene specifying the disease. For this reason, establishing linkage of a random polymorphic probe with a disease locus requires examination of the DNA from both parents and the affected offspring, in order to determine the "phasing" of the polymorphism with the disease in each family.

The use of RFLP mapping has been very powerful for establishing the segregation of several disease loci with particular polymorphic probes. These "genetic mapping" experiments become most valuable when a genetic map is linked up with a physical map; that is, when a disease is found to be genetically linked to a polymorphic probe (which, remember, is a cloned piece of DNA), one can then determine the physical location of the probe and therefore then have some idea of the physical location of the disease gene. If linkage is found, the first experiment that is usually done is to determine on which chromosome the probe resides, usually by in situ hybridization or Southern blotting to DNA from somatic cell hybrids carrying a single human chromosome (the localization of the probe to a particular chromosome is a low-resolution physical map). Then a more precise localization of the probe can be made by using the approaches outlined in this report.

A disease locus gene can eventually be cloned by using this approach, which has been referred to as "reverse genetics" - the isolation of a gene on the basis of its chromosomal location without regard to its biochemical properties. This approach has led to the isolation of the genes causing chronic granulomatous disease (Royer-Pokora et al. 1986), Duchenne muscular dystrophy (Monaco et al., 1986), retinoblastoma, and probably cystic fibrosis (Estivill et al., 1987); many more genes are likely to follow. The first step is to establish linkage of the disease with a cloned polymorphic marker. Then the region around the marker is mapped physically, leading to the isolation of candidate genes in the region. Once candidate genes are isolated, their expression in various tissues, at various developmental stages, and in individuals with and without the disease is tested. These approaches can lead to the identification of an expressed sequence that is very likely the disease gene; positive identification is obtained when the locations of mutations causing the disease phenotype are determined, and, ultimately, when the

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

cloned sequence is shown functionally to cause the phenotype when altered.

One technical difficulty with reverse genetics is the need to find highly polymorphic probes. From the previous discussion, it is clear that the more polymorphic a probe is, the more valuable it is for linkage studies. A problem with the RFLP approach is that many single base changes do not create or destroy a restriction enzyme cleavage site, so they are not detected by this procedure [the chance that a particular single base change will be detected by a given restriction enzyme is less than 0.5%]. Thus, even when many different restriction enzymes are tested with a particular probe, there is a good chance that many polymorphisms will be missed; in fact, many probes that are isolated eventually end up being useless because little or no polymorphism can be detected by restriction enzymes. For this reason, alternative approaches to detect single base changes in genomic DNA have been developed. Two methods were were recently developed, the first is called denaturing gradient gel electrophoresis (Fischer and Lerman, 1983; Myers et al., 1985a), and the second is called RNase cleavage (Myers et al., 1985b). The details of both of these methods are reviewed in Myers and Maniatis, 1986. The feature of these approaches is that they both allow a much larger fraction of all possible single base changes in a genomic DNA sample to be detected; it is estimated that over 60% of all possible single base changes can be detected by each of these methods. Indeed, other investigators have used these approaches to detect mutations and polymorphisms that were not detected by restriction enzymes (Chehab et al., 1986; Noll and Collins, 1987; B. Levinson and J. Gitschier, personal communication).

OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

Appendix 3: CALCULATION OF NUMBERS OF COSMID CLONES NEEDED TO REPRESENT THE ENTIRE GENOME

The following arguments were obtained from Clarke and Carbon (1976) and Maniatis et al. (1982). In order to determine how many cosmid clones are necessary to cover the entire human genome (or a designated fraction of it), the following formula is used:

$$N = \ln (1-P)/\ln (1-f)$$

where P is the designated fraction of the genome (also can be stated as the designated probability of finding any particular DNA sequence in the library), f is the fractional proportion of the genome present in a single cosmid, and N is the necessary number of cosmid clones to achieve the designated probability. For example, to achieve a 99% complete library (or to have a P = 0.99 of containing a particular DNA fragment) with cosmid inserts of 40,000 bp of the human genome (3×10^9 bp):

$$N = \ln (1 - 0.99) / \ln (1 - [4 \times 10^4/3 \times 10^9])$$

$$= 346,251$$

Since the average cosmid insert is 40,000 bp and the human genome is 3×10^9 bp, about 75,000 exactly contiguous cosmid inserts would represent the entire genome. Therefore, 346,251/75,000 = 4.6; the number of cosmids needed to ensure 99% completeness represents about 5-fold redundancy of the sequences in the genome. Note that this result holds for any size insert in any cloning system; five-fold redundancy in the yeast cloning system would also result in 99% complete artificial chromosome libraries.

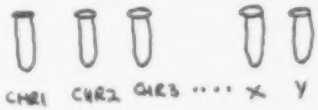
OTA DRAFT: DO NOT CITE, QUOTE, OR REPRODUCE WITHOUT PERMISSION

FIGURE 1

SCHEME FOR MAPPING THE HUMAN GENOME

STEP

1 ISOLATE EACH HUMAN CHROMOSOME — FLOW CYTOMETRY AND SOMATIC CELL HYBRIDS



PURIFIED HUMAN CHROMOSOMES

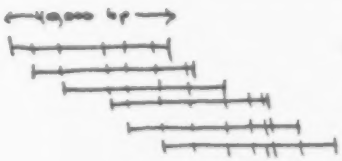
2 DIVIDE EACH CHROMOSOME INTO 50 – 200 FRAGMENTS (0.5–5 MILLION BP IN LENGTH) — YEAST ARTIFICIAL CHROMOSOME CLONING SYSTEM

YEAST ARTIFICIAL CHROMOSOME LIBRARIES: 100–1,000 YEAST CLONES CARRYING OVERLAPPING 0.5–5 Mbp HUMAN DNA FRAGMENTS

3 SUBDIVIDE EACH 0.5–5 Mbp SUBCHROMOSOMAL FRAGMENT INTO OVERLAPPING 40,000 bp FRAGMENTS — MAKE COSMID LIBRARIES FROM EACH YEAST ARTIFICIAL CHROMOSOME

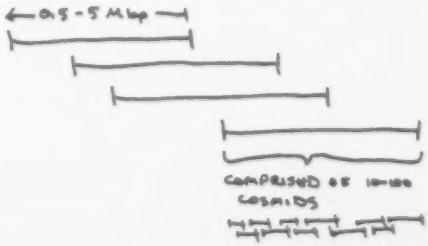
50–500 COSMID CLONES CARRYING OVERLAPPING HUMAN INSERTS

4 DETERMINE ORDER OF RESTRICTION SITES WITHIN EACH COSMID & USE THIS INFORMATION TO DETERMINE THE ORDER OF THE COSMIDS RELATIVE TO EACH OTHER — DIGEST EACH COSMID DNA WITH A RESTRICTION ENZYME, EXAMINE PATTERNS BY GEL ELECTROPHORESIS



CONTIG MAP OF COSMIDS

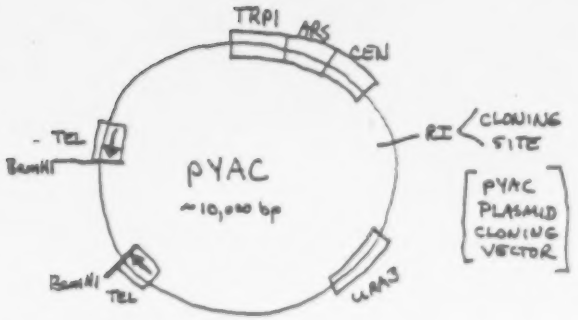
5 LINK TOGETHER 0.5–5 Mbp DNA FRAGMENTS INTO INTACT CHROMOSOMES — USE COSMID OVERLAP INFORMATION TO CONNECT 0.5–5 Mbp FRAGMENTS



CONTIG MAP OF YEAST ARTIFICIAL CHROMOSOME

FIGURE 2

YEAST ARTIFICIAL CHROMOSOME CLONING SYSTEM
[FROM BURKE, CARLE, & OLSON, 1987]



[pYAC PLASMID CLONING VECTOR]

CAREFULLY-PREPARED GENOMIC DNA

TEL = yeast telomere sequences

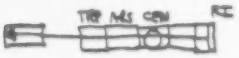
CEN = yeast centromere sequences

TRP1, URA3 > YEAST SELECTABLE MARKERS

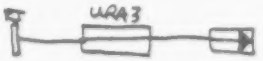
ARS = autonomous replication sequences

DIGEST WITH BAMHI, ECORI & PHOSPHATASE

PARTIAL ECORI RESTRICTION ENZYME TREATMENT



LEFT ARM

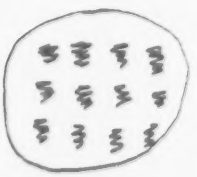


RIGHT ARM

0.5-5 MB DNA FRAGMENTS

T4 DNA LIGASE

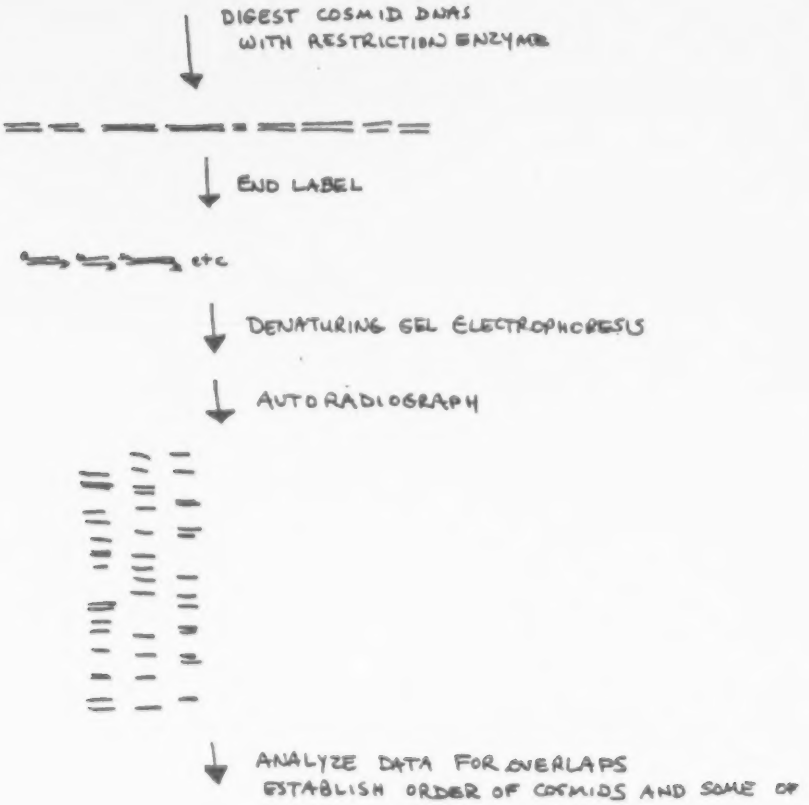
TRANSFORM DNA INTO YEAST
SELECT FOR TRP1 & URA3

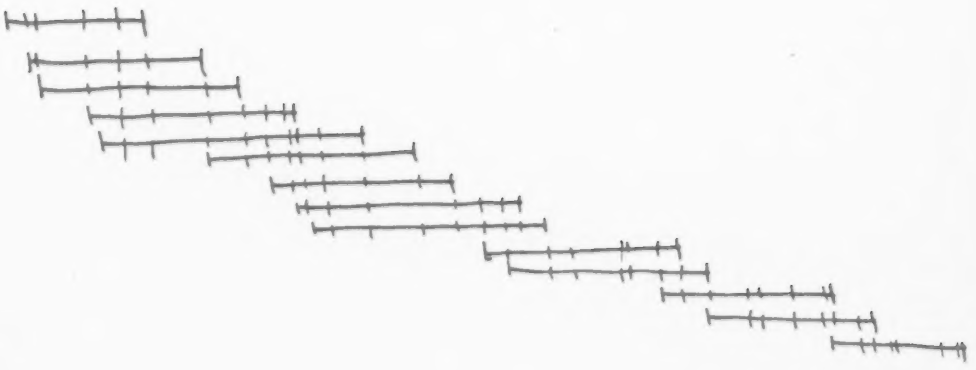


YEAST STRAINS CARRYING ARTIFICIAL CHROMOSOMES

FIGURE 3

CONTIG MAPPING OF COSMIDS AND DETERMINATION OF THE ORDER OF RESTRICTION SITES


DIGEST COSMID DNAS WITH RESTRICTION ENZYME
END LABEL
etc
DENATURING GEL ELECTROPHORESIS
AUTORADIOGRAPH
ANALYZE DATA FOR OVERLAPS ESTABLISH ORDER OF COSMIDS AND SOME OF THE RESTRICTION SITES



23 Streatham Common South
London SW16 3BX
England

12 October 1987

Dear Ms Courteau,

I enclose:
1. My finalized report "Mapping and Sequencing the Human Genome in Europe".
2. A disk holding the report in separate files corresponding to the report's chapters: file names specified on disk holder. The files were written using Wordstar on an IBM compatible PC.
3. Copies of two relevant documents.

In rewriting my report I have taken into account just about everything you raised for which information was available. Note in particular:
1. There was no follow-up EMBO meeting in September.
2. The June meeting between Brenner, Tooze, Philipson and possibly others of which Bob Cook-Deegan had heard is denied (or forgotten) by Brenner, who thinks this must refer to the original EMBO meeting.
3. Rechaussat of CNRS heads a small internal advisory group.
4. The French government have just decided to fund a small programme in human genome research but details are not yet available.
5. A new French database, GENATLAS, has recently been announced.
6. Details of the CNR project have been added and the document is enclosed.
7. A few more details of the Brenner and CEPH projects have been added but full details do not exist.
8. Time being short and it not being essential, I have not added a conclusion to the report.

If there is anything I have forgotten, please let me know. I hope the report proves useful. Am I permitted to distribute copies to those who have helped me most?

Yours sincerely

Peter Newmark

MAPPING AND SEQUENCING THE HUMAN GENOME IN EUROPE

Peter Newmark D.Phil, B.A.
Nature, London

Prepared for and submitted to OTA on 13 October 1987.

CONTENTS

1 EUROPEAN ORGANIZATIONS

1.1 Context
- 1.1.1 European Community
- 1.1.2 European Molecular Biology Organization
- 1.1.3 European Molecular Biology Laboratory
- 1.1.4 European Science Foundation
- 1.1.5 European Research Coordination Agency

1.2 Work in Progress and Planned
- 1.2.1 European Community
- 1.2.2 European Molecular Biology Organization
- 1.2.3 European Molecular Biology Laboratory
- 1.2.4 European Science foundation
- 1.2.5 European Research Coordination Agency

1.3 References

2 FRANCE

2.1 Context

2.2 Work in Progress and Planned
- 2.2.1 Centre d'Etude Polymorphisme Humain
- 2.2.2 CNRS/INSERM/Ministry of Research
- 2.2.3 Pasteur Institute

2.3 Collaboration

2.4 References

3 FEDERAL REPUBLIC OF GERMANY

3.1 Context

3.2 Work in Progress and Planned

3.3 References

4 ITALY

4.1 Context

4.2 Work in Progress and Planned

4.3 References

5 UNITED KINGDOM

5.1 Context

5.2 Work in Progess and Planned
5.2.1 Medical Research Council
5.2.2 Imperial Cancer Research Fund
5.2.3 Science and Engineering Research Council
5.2.4 Other Work
5.2.5 Departments

5.3 Collaboration

5.4 References

6 OTHER COUNTRIES

7 DATABASES

7.1 Context

7.2 Existing databases
7.2.1 International
7.2.2 National

7.3 Plans

7.4 References

7.5 Database Addresses

8 REFERENCES

1 EUROPEAN ORGANIZATIONS

1.1 CONTEXT

1.1.1 European Community

A framework for Research and Development within the European Community dates back only to 1974 and, because of the Community's origins in coal, steel and nuclear research, the framework at first only accommodated biological research in the context of health hazards. From the mid 1970s the Commission of the European Communities began proposing the support of biotechnological research and by 1981 the Council agreed to accept such a proposal, despite the lack of any formal provsion within the founding Treaty of Rome for research that was not directly related to the economies of the community. The Single European Act that has replaced the Treaty of Rome does include provsion for such research, which is to be precompetitve in nature and is to be aimed "to strengthen the scientific and technological basis of European industry and to encourage it to become more competitive at international level".

Following the Council decision in 1981, the four-year Biotechnology Engineering Programme was launched, followed by the Biotechnology Action Programme (BAP)

which runs until the end of 1989, when it will be replaced by a programme that is provisionally known as Biotechnology Research for Innovative Development and Growth in Europe (BRIDGE). When the $55 million BAP was approved, it was foreseen that revisions might be necessary to increase the programme and the budget. Plans for revision had to be postponed when the Council failed to reach agreement on time on the budget for the framework of overall research and development for the period 1987-1991. A limited agreement was finally reached mid way through 1987 and some $23 million is likely to become available for revisions of BAP with $115 million for BRIDGE (including training and concertation). Within both of these programmes there is some intention to fund work of relevance to human genome mapping/sequencing. There is also likely to be relevant research within the health budget.

The money will be spent on areas of research that are proposed by the Commission and approved by Council and the European Parliament. The Commission's choice is guided by consultation and advice. It issues discussion papers and a summary of responses to them, as well as contract reports that are also published. A Concertation Unit for Biotechnology in Europe (CUBE) has been in existence since 1983 to help the commission as a whole to monitor worldwide

developments in biotechndlogy and, in particular, to coordinate the biotechnological activities of the separate sectors of the commission (especially sectors DG XII, DG VI and DG XIII, which are concerned with science R & D, agriculture, and information market and innovation, respectively). CUBE has played no part in overseeing or coordinating funded commission projects so far and would be unlikely to play such a part in any genome mapping project.

After the Biotechnology Action Programme had been approved, there was an open call for proposals with priority being given to transnational proposals that included an industrial partner or at least provided evidence of specific industrial interest. There is likely to be even more emphasis on industrial participation in BRIDGE. Intellectual property rights on any discovery financed by European Community funds belong to the contractor. If the contractor does not wish to exercise his patent rights, the Community has the second option.

1.1.2 European Molecular Biology Organization (EMBO)

EMBO (which is housed in the European Molecular Biology Laboratory - see 1.1.3) is funded by most European countries to fulfil a role in the training of European molecular biologists. The great majority of

its money is spent in supporting fellowships that allow European researchers to spend time in laboratories other than their own, generally within Europe. It also supports workshops and training courses, a few meetings and a journal. But it does not support research itself.

1.1.3 European Molecular Biology Laboratory (EMBL)

This laboratory in Heidelberg is financed by contributions from all the major and most of the minor countries of Europe. EMBL houses the administrative offices of EMBO and EMBO's Secretary, John Tooze, serves as part-time Scientific Coordinator of EMBL, but EMBO and EMBL have completely separate budgets and some countries that contribute to EMBO do not contribute to EMBL.

EMBL has a budget of about DM45 million and a staff of about 250, of which the scientists are drawn from all member nations (and a few from other countries) whereas the technical staff are mostly German. EMBL and certain FRG national laboratories in Heidelberg (the German Cancer Research Center, a Max Planck Institute, and the University of Heidelberg's Center for Molecular Biology have had a mutually benefical effect, with the result that the city is the main focus for FRG molecular biology/biotechnology.

EMBL's scientific programme is subject in outline to the approval of the Council, on which each contributing country is represented, and is proposed by the laboratory's Director-General, currently Lennart Philipson. The laboratory was conceived on the notion that molecular biology would require facilities that would be too expensive for any national research programme to support. By and large this has turned out to be a misconception and EMBL has been unable to fulfill its supposed transnational function. Instead it has developed into a good, but not exceptional, laboratory which attracts the criticism from certain countries (notably the UK) that it does little that cannot be done - and done more cheaply - at the national level. Such countries have at times been grudging in their support of the laboratory.

1.1.4 European Science Foundation

This organization, with administrative headquarters in Strasbourg, is subscribed to by research councils and equivalent bodies from European countries. It has a very small central fund but increasingly supports projects on a special funding basis. It has tended to steer clear of biology but has recently supported some protein engineering work. It has a standing committee called the European Medical Research Councils that enables the heads of national medical research bodies

to meet once a year. But this committee has no budget and if it has much influence it is only within the foundation.

1.1.5 European Research Coordination Agency (EUREKA)

A French-initiated response to the US Strategic Defence Initiative, EUREKA was set up in 1985 to encourage the development of advanced technologies within Europe. It promotes industry-led technological collaboration between European countries in areas that include biotechnology and advanced information technology. It is not intended as an alternative to the efforts of the European Community, whose role is seen to be less close to the market. A EUREKA project must involve at least two industrial laboratories in two different European countries. Governments vary in the extent to which they provide financial support for EUREKA projects. Some tend to offer no more than token support - oiling the wheels and helping to sort out any problems of international collaboration. Others have provided real finance on occasion and the French government will pay 50% of a EUREKA project. A small secretariat in Brussels coordinates the activities of EUREKA.

1.2 WORK IN PROGRESS AND PLANNED

1.2.1 European Community

Biotechnology Action Programme

Within the subprogramme on Contextual Measures there is one contract for Automation of DNA Sequencing and several concerned with data capture, processing and banking (Commission.....XII/87-BN/hl).

The Automation of DNA Sequencing contract is to the University of Manchester Institute of Science and Technology (UK) the University of Konstanz (FRG) the Rutherford Appleton Laboratory (UK) and the Transfer Centre for Image Data Processing, Konstanz (FRG). The Manchester group has already developed an automatic reagent manipulating system (Martin) and together with the Transfer Centre is developing instrumentation for automated plaque selection and culture inoculation (Martin and Davies). Apparatus for automated direct-blotting electrophoresis has been developed at the University of Konstanz (Beck and Pohl 1984) and the Rutherford Appleton Laboratory is working on multi-wire proportional counters that will scan either the output of the apparatus or standard sequencing gels for 35-sulfur and digitize the band pattern (Martin and Davies). (For an overview of the project see

Martin and Davies, 1986) There is no immediate prospect of linking the components together but separate instruments may be brought to the market. Thus the automatic reagent manipulating system has been licensed to Beckman Instruments Inc. and Hoefer Scientific Instruments (in California) are commercializing the automatic electrophoresis apparatus (Marin and Davies). No immediate plans exist to employ these instruments in human genome sequencing in part because they need further devlopment and in part because the natural interests of the Manchester group lie with the yeast genome (Oliver).

Other BAP contracts with some relevance to human genome mapping and sequencing are titled as follows:

* Electrophoresis of proteins: Data capture, analysis and construction of databanks (Belgium, UK and France)
* Computer-aided peptide and protein engineering software development (UK and France)
* Research and diffusion for a portable European system of access and analysis of biosequences (France and Ireland)
* Protein sequence data bank (FRG) (see 7.2.1)
* Artifical intelligence approach to protein structure prediction by development of a knowledge base (Belgium, FRG, and EMBL)

The first of these contracts involves Joyce-Loebl as

an industrial partner and the second involves the Wellcome Foundation. None of the others has an industrial partner.

BRIDGE

While not yet finalized, this programme is likely to include a focus on sequencing the genome of yeast, which is seen as more feasible than sequencing the human genome and of considerable assistance towards that ultimate goal (Goffeaut). As conceived, the plan would be to start with perhaps 15 laboratories, one yeast chromosome and a 2-year pilot project but eventually to involve a large number of European yeast laboratories. A pilot project, consisting of four small studies in the first place, will be included in the revisions to the Biotechnology Action Programme. The full project would be part of BRIDGE and is provisionally costed at $50 million. The project, which has had mixed reactions, would be intended in part to create a market for sequencing equipment (Cantley).

Bio-informatics

A Task Force for Biotechnology Information was created in 1982 by the Commission's Directorate-General for information market and innovation (DG XIII). It reports to DG XIII's Committee on Information and Documentation in Science and Technology, and has CUBE

as its scientific secretariat. The task force has representatives from all member states and has a mandate that emphasizes the development of databases and information services for the support of biotechnology. Its activities are closely coordinated with BAP's Contextual Measures subprogramme and will presumably be similarly involved with BRIDGE. The Task Force has not only produced discussion papers but has also provided small sums of money, totalling $200,000 to EMBL's nucleotide sequence data library (see 1.2.3), to a workshop on nucleotide sequences, software and user needs (Commission...XII.604.86) and to databanks for hybridomas and microbial resources. Its workplan for 1986-1988 (Commission...BIOT/5/86) envisages more support for EMBL, an international protein sequence bank and the hybridoma bank. Regarding EMBL, one possibility might be to support its expanding biocomputing programme (see 1.2.3), another could be to network the EMBL nucleotide sequence database (Nieuwenhuis).

In 1986 the task force together with CUBE initiated a planning exercise known as Bioinformatics: Collaborative European Programmes and Strategy (BICEPS), which "aims to formulate a mid to long-term strategy for Europe in bio- and medical informatics" and "overall, to improve the European competitve position in the rapidly developing world market for these technologies and applications" is investigating

the interface between biotechnology and information technology (Commission...XII/86). There are references in documents on bioinformatics generated by BICEPS (for example, Flensholt et al., Coulson) to human genome sequencing, and there is no doubt that some of these will turn into firm proposals, such as DG X111's AIM (Advanced Informatics in Medicine).

Programme of Predictive Medicine and Novel Therapy (PROMPT)

This proposal, which falls within the health sector of the framework for research and development 1987-1991, proposes supporting research in four areas of research: study of the human genome, nucleic acid probes, genetic therapy and monoclonal antibodies. The first of these is largely related to mapping the human genome as an aid to the diagnosis and prevention of genetic disease. Originally proposed as a $75 million programme, PROMPT has been revised downwards to $25 million. Both the budget and the content of PROMPT may be further revised before approval.

1.2.2 European Molecular Biology Organization

A meeting of Europeans with an interest in human genome mapping/sequencing was sponsored by EMBO in Paris in February 1987. Little interest was expressed in mounting a major European project. Most participants seemed to favour informal co-operation

between individual laboratories. There was pessimism concerning the prospect of any large-scale public funding in Europe; the prospect of raising money from private sources was discussed (Tooze). A planned follow-up meeting in September did not take place, apparently for lack of interest.

1.2.3 European Molecular Biology Laboratory

Relevant programmes of research (EMBL 1985) at EMBL are:

* Instrumentation: development of techniques for automated DNA sequencing under W. Ansorge; includes work leading to replacment of film autoradiography by electronic detection, to automation of the DNA sequencing technique with radioactive and fluorescent labels and to automation of the reactions and gel loading. EMBL holds patents. Technology has been licensed to LKB (Sweden) for commercialization. Talks between EMBL and possible European collaborators are likely to result in a joint project to sequence a megabase (one million bases) of human DNA (Philipson).

* Biocomputing: the EMBL Data Library shares with the GenBank in the United States the task of collecting, organizing and distributing a nucleotide sequence database (see 7.2.1). A recently approved expansion of the biocomputing programme foresees further development of this database and its linkage

to, for example, the relational database at Birkbeck College, London, the protein engineering programme in Groningen and the computing center at the Pasteur Institute. The expanded biocomputing programme will include a group devoted to the development of sequence analysis techniques (EMBL 1986).

* Gene mapping/sequencing: advanced techniques for mapping have been under development but the group involved, led by Hans Lehrach, has now left EMBL.

For EMBL to become seriously involved in human genome mapping/sequencing, as its director-general would like it to be, there would need to be considerable enthusiasm from the contributing nations, which seems to be lacking (Philipson, Tooze), in part because of the perceived failings of the EMBL Data Library (see 7.2.1). There would also need to be considerable additional funding of the laboratory, which seems most unlikely in present circumstances. In 1986 EMBL's director-general wrote:

"If the American plan to launch a programme on the human genome materializes the EMBL may be a natural collaborative partner in this project. It might, therefore, be worthwhile to plan for at least one new Programme in one of those fields to be initiated in Heidelberg at the end of the proposed Scientific Programme (1990). To facilitate recruitment and the

launching of this Programme, plans should be available by 1990 but we do not foresee any cost during the next 4 years." (EMBL 1986).

1.2.4 European Science Foundation

The foundation's standing committee, the European Medical Research Councils, decided at its 1987 meeting not to attempt to coordinate European research on human genome mapping/sequencing.

1.2.5 European Research Coordination Agency

EUREKA may play a part in linking French research on human genome mapping/sequencing to industrial partners in Europe (see 2.3).

1.3 REFERENCES

Beck,S. and Pohl,F.M., "DNA sequencing with direct blotting electrophoresis" EMBO J. 3: 2905-2909,1984.

Cohen, D., Staff, Centre d'Etude du' Polymorphisme Humain, Paris, France, personal communication, July 1987.

Commission of the European Communities, Biotech Info in the EEC Work Plan 1986-87-88, BIOT/5/86, Brussels, 1 October, 1986.

Commission of the European Communities, Biosequences: Perspectives and User Services in Europe,

XII.604.86, Brussels, 1986.

Commission of the European Communities, Bio-informatics: Collaborative European Programmes and Strategy, draft XII/86, Brussels, 21 April 1986 (with revised Annex I-30.9.86).

Commission of the European Communities, List of Research Contracts in the Biotechnology Action Programme, XII/87-BN/hl, Brussels, 7 June 1987.

Coulson,A.F.W. and Collins,J.F. "Biological Sequence Analysis: Advanced Architecture Computing Requirements", XII/604/86, prepared for the Commission of European Communities, Brussels, June 1986.

EMBL Annual Report 1985

EMBL Annual Report 1986

EMBL, The 1987-1990 Scientific Programme of the EMBL, EMBL/86/3 rev. 1 E, Heidelberg, 1 September 1986.

Flensholt, J., Jacobsen, H.K. & Jakobsen, L.K. "Proposal for a European Action Plan in Bio-Informatics" draft report prepared for the Commission of the European Communities BICEPS, Brussels, December 1986.

Goffeaut, A., DG XII, Commission of the European Communities, Brussels, personal communication, July 1987.

Martin,W.J., Warmington,J.R., Galinski,B.R. et al. "Automation of DNA Sequencing" Bio/technology 3:911-

915, 1985.

Martin, W. J. & Davies, R.W., "Automated DNA Sequencing: Progress and Prospects" Bio/technology 4: 890-895, October 1986.

Nieuwenhuis, B., DGXII Commission of the European Communities, Brussels, personal communication, 7 July and 8 October 1987.

Oliver, S., Director, Manchester Biotechnology Centre, University of Manchester Institute of Science and Technology, Manchester, UK, personal communication, July 1987.

Philipson,L., Director-General, European Molecular Biology Laboratory, Heidelberg, FRG, personal communication, July 1987.

Tooze,J., Secretary, European Molecular Biology Organization, Heidelberg, FRG, personal communications, June and July 1987.

2 FRANCE

2.1 CONTEXT

Science in France is probably more politicized and more driven by nationalism than in any other country in Europe. From 1981, declaring his aim to make France the third scientific and technological power of the world, Prime Minister (now President) Mitterand set about delivering both more money for science and some political interest in its progress. Under his government, moreover, many French academics were persuaded for the first time to cooperate with industry. That cooperation is also favoured by the new government that took over in 1986 even though many of the other reforms brought about under Mitterand are not to its liking. French scientists are therefore going through a further period of change.

Government funded research is mostly channelled through the research organizations which come under the wing of the Ministry of Research and Higher Education. Almost all research is carried out within universities, often in units set up by the research organizations. The largest organization of all is the Centre National de la Recherche Scientifique (CNRS). It, and the much smaller Institut National de la Sante et de la Recherche Medicale (INSERM), are the only two organizations that fund research of direct

relevance to human genome mapping/sequencing. A separate centre of relevant research is the Pasteur Institute in Paris, which is a semi-automonous institute that receives about half its funds from the state.

But the real focus of attention is the Centre d'Etude du Polymorphisme Humain (CEPH), founded by Jean Dausset and currently housed in the College de France in Paris. With Dausset's impending retirement from the College de France, CEPH will be moved to the Hopital Saint-Louis, also in Paris. CEPH is a non-profit, independent organization that draws the bulk of its financial support from private sources in the form of a permanent endowment, with some additional funds from INSERM. As its name implies, CEPH was set up in the first place to :

"hasten the mapping of the human genome by linkage analysis with DNA polymorphisms. A key premise of the CEPH is that the genetic linkage map will be efficiently achieved by collaborative research on DNA from the <u>same</u> sample of human families. To this end, the CEPH provides to collaborating investigators high quality, cellular DNA produced from.....each member of a reference panel of large families and a data base contributed to and shared by the investigators. Collaborating investigators use the DNA from the CEPH panel to test the families for segregation of

restriction fragment length polymorphisms detected with their oen probes. they contribute the results of these tests to the data base." (Dausset).

Part of CEPH's self-appointed role is to cevelop software for data preparation and linkage analysis to help both collaborating investigators and the analysis of the large quantity of data entering the central database. There is both a CEPH Collaborative Data Base, which began this year, contains both published and unpublished data and is available only to collaborators, and a CEPH Public Data Base, which will be activated and periodically updated with only the published data from the Collaborative Data Base after a "suitable" lag period, whereupon it will be made available to the scientific community at large.

2.2 WORK IN PROGRESS AND PLANNED

2.2.1 Centre d'Etude Polymorphisme Humain

The original aim of CEPH - to produce a 20 centimorgan linkage map of the complete human genome - will be completed during 1988. The 150-200 restriction fragment length polymorphism probes that define the map will then be distributed, free of charge, to any interested party on condition that they are not used for commercial purposes (Cohen).

A new aim of CEPH is to become involved in large-scale

sequencing of DNA and the analysis of sequence data (Cohen). As a start, the plan is to sequence 3 megabases of DNA from the human HLA locus. Michael Steinmetz, formerly of the Basel Institute of Immunology but now with Roche in Basel, is well advanced in preparing the necessary cosmids from a homozygous individual from a CEPH family. Sequencing, which has already begun, will be assisted by two prototype sequenators, one from Applied Biosystems, the other from the Swiss company Genofit.

Jean Dausset, CEPH's director is exploring the possibility of raising political and financial support for the expansion of the center's activities into mapping/sequencing. If he is successful, several researchers with interests in specific chromosomal regions or in methods of data analysis may become involved in the project (Cohen).

2.2.2 CNRS/INSERM/Ministry of Research and Higher Education

None of these organizations, has declared a plan for human genome mapping/sequencing but each is considering what part it might play (Chambon, Thomas). The French research budget, declared in the fall of 1987, includes a National Programme of Research into the human genome. It is thought likely to be budgeted

at only about 10 million francs, probably to be distributed through the standard grant mechanism. Within INSERM there is a group of three people, headed by Louis Rausellat, whose role is to define objectives and offer advice to the director on human genome mapping/sequence (Rechaussat).

There are, of course, a number of individual scientists working in relevant areas who are funded by CNRS or INSERM and who would be strong candidates for being involved in some larger-scale effort. One example is Jean-Louis Mandel of the University of Strasbourg who works on the X chromosome (Mandel).

2.2.3 Pasteur Institute

No specific project is planned at the Pasteur but J.M. Claverie has a major interest in writing programmes to make sense of large sequences (Claverie). His laboratory holds one of the European Commission's contracts for An Artifical Intelligence Approach to Protein Structure Prediction by Development of a Knowledge Base (see 1.2.1).

2.3 COLLABORATION

No collaborations yet exist in France but there is a strong undercurrent of opinion that favours a substantial programme in human genome mapping/sequencing as long as it is not funded at the

expense of other research. It is widely believed that human genome work could be as important as a spur to technological development as for its scientific output.

In respect of technological developments, the intention is to try and work through EUREKA (see 1.1.5) to involve European companies with an interest in instrumentation or information technology. The French government (usually through its ministry of industry) is prepared to provide 50% funding for EUREKA projects and there are indications that it would consider CEPH's human genome work as a valid case for EUREKA funding (Cohen, Thomas).

On the scientific side, there is some optimism that additional private and, perhaps, state funding will be found to expand the activities of CEPH in conjunction with individuals currently employed by CNRS, INSERM and the Pasteur Institute (Cohen).

2.4 REFERENCES

Chambon,P., Faculty of Medicine, University of Strasbourg, personal communication, July 1987.

Claverie,J.-M., Institut Pasteur, Paris, personal communication, July 1987.

Cohen,D., CEPH, College de France, Paris, personal communications, July and October 1987.

Dausset,J., Cann,H. and Cohen,D. "Centre d'Etude du Polymorphisme Humain (CEPH): Collaborative Mapping of the Human Genome" background paper CEPH, College de France, Paris , undated.

Mandel,J.-L., Faculty of Medicine, University of Medicine, personal communication, July 1987.

Rechaussat,D. INSERM, personal communication, October 1987.

Thomas,D., Technical University of Compiegne, personal communication, July 1987.

3 FEDERAL REPUBLIC OF GERMANY

3.1 CONTEXT

Molecular genetics is less strong than it might be in FRG for three reasons. First, the revolution in molecular biology occured when the country was still suffering from the emigration of many of its scientists and the effects of World War 2. Second, because of the abuse of genetics during Hitler's reign, genetics has continued to have unfortunate connotations in the German context. Third, a strong 'Green' movement, opposed to genetic manipulation, has emerged in FRG, making it difficult to gain political support for molecular genetic research. The federal government has before it a report from a specialist committee that recommends a moratorium on genetic research on human embryos and is considerably restrictive in other respects (Deutscher Bundestag). The current climate of opinion is therefore considered to be unfavourable to any large-scale work on the human genome, although it has not prevented the development in Koln of the strongest base of plant molecular genetics/manipulation in mainland Europe.

Research funding in FRG, as in most of Europe, is at a standstill. There are four main sources of funds. First, the Max Planck Society which supports the Max

Planck Institutes, each of which is devoted to a particular, albeit broad, area of research, such as Biochemistry. Second, the German Research Association (Deutsche Forschungsgemeinschaft) which recieves approximately half of its funds from the federal government and half from state governments, and which supports research in universities. Third, funds are available through the German Ministry for Research and Technology (Bundes Ministerium fur Forschungs und Technologie), which, for example, supports the Institute for Biotechnology Research in Braunschweig, but also supports projects in universities. And, fourth, there is science supported at state level only.

3.2 WORK IN PROGRESS OR PLANNED

There is no current project directly aimed at large-scale human genome mapping/sequencing within FRG and none seems likely to emerge in the immediate future for the political and financial reasons already mentioned.

The German Research Association, which would almost certainly have to be the source of public funds is in the process of assessing its position and has made no decions. At a recent closed meeting of interested parties, however, little prospect emerged of any

programme of research being mounted. The Max Planck Society neither has an appropriate institute nor the available funds to initiate a relevant research programme.

There are, however, several laboratories that are already carrying out relevant research that could be extended at little expense and linked with an international collaboration. For example, current work on sequencing the locus of the immunoglobulin kappa light chain and the loci at which the hepatitis B virus integrates into the human genome could be extended and fed into an international database. There is an emerging view that any such database should be in the public domain. It is also thought possible that work on the preparation and sequencing of tissue-specific cDNA libraries might begin in FRG.

There are two European Commission contracts in FRG that are relevant to human genome mapping/sequencing. Fritz Pohl at the University of Konstanz is one of the holders of a contract for the automation of DNA sequencing (see 1.2.1) and D. Oesterhelt at Max Planck Institute for Biochemistry has a contract for a protein sequence data bank (see 7.2.1) (Buschbeck, Hofschneider, Zachau).

3.3 REFERENCES

Buschbeck, F., Bundes Ministerium fur Forschungs und Technologie, Bonn, personal communication, July 1987.

Deutscher Bundestag "Chancen und Risiken der Gentechnologie",d.Bericht d. Enquete-Kommission d. 10 Deutscher Bundestages, Deutsches Bundestag, Referat Offentlichkeitsarbeit Bonn, 1987.

Hofschneider,P., Director, Max Planck Institut fur Biochemie, Martinsreid, personal communication, 14 July 1987.

Zachau,H., University of Munich, personal communication, 26 June 1987.

4 ITALY

4.1 CONTEXT

While Italian governments come and go at an alarming rate, recent versions have all given priority to improving the country's scientific performance with the intention of sparking a technology-led revitalization of Italy's economy. Considerable extra funds for technology-related research have been available in the past few years, with biotechnology as one focus. Italian biotechnology and molecular biology start from a low base (Vezzoni) and can hardly fail to improve. The question is whether the available funds and the evident enthusiasm can be converted into substantial progress. It would be fair to say that outside Italy there is little tendency to take the Italian project seriously and even some of those involved in the project are somewhat sceptical of its chances of success.

Government funds for the more applied end of biotechnology are distributed as research contracts, largely either to the universities or to the Consiglio Nazionale delle Richerche (CNR), Italy's national research council. CNR, itself, is responsible for

supporting almost all the more academic end of biotechnology/molecular biology research both in universities and in its own research institutes

4.2 WORK IN PROGRESS AND PLANNED

In June 1987, the CNR announced its decision to fund a project entitled "Mapping and Sequencing the Human Genome" (Mappaggio e Sequenziamento del Genoma Umano). A combination of advantages provided the justification, according to the CNR.

First, the project could provide a means rapidly to increase Italian activity in molecular biology. Second, it should stimulate the acquisition and development of instumentation and techiques that are important for advances in biotechology. Third, because the project would yield large quantities of data, it would encourage the introduction of new methods of data analysis - an area that is due to develop rapidly. Finally, since international cooperation in human genome mapping/sequencing is likely to evolve. Italian invelvment would give young researchers access to the sophisticated technology of more advanced countries (CNR document).

In the first place, CNR has provided just over $1 million a year for a two-year 'strategic project', the outcome of which will determine whether a larger-scale 'finalized project' is funded at around $10 million a

year (note that these sums are to cover only specific materials, machines, travel, meetings etc. and not salaries and general overheads since only existing personnel will be involved) (Vezzoni). The most direct aim of the strategic project is to test how feasible it would be to split a single chromosome into a number of fragments whose relationship to one another would be ascertainable by cytological means. If so, it would represent a considerable short cut to the more conventional route to a convenient starting point for cloning and sequencing (Dulbecco).

A key question to be answered in the strategic project is whether it is possible to isolate a single chromosome without damaging it to the extent that it would be futile to undertake an attempt to sequence it. If it does seem possible to start with a single chromosome, the question will be which one, given that it should be as small as possible but also 'interesting' (Dulbecco). The choice would probably be between chromosomes 21,22 or Y. If the preferred strategy is not possible, consideration will have to be given to a project based on the more conventional technique of sequencing random and cDNA-identified phage or cosmid clones. The possibility of using yeast megacloning vectors (Burke) will also be explored.

The strategic project, which is to be coordinated by

Renato Dulbecco of the Salk Institute in California who will pay frequent visits to Italy, will fund the following (CNR) :

* cytogenic analysis and in situ hybridization (A.Forabosco, University of Modena; P.Benedetti, CNR Pisa).
* chromosome sorting (N.M.Maraldi, CNR Bologna; G.Bottiroli, CNR Pavia).
* megaclones, cDNA and chromosome libraries, large scale mapping (P. Vezzoni and D. Brevario, CNR Milano; G. Tocchini Valentini, CNR Rome; D'Urso Boncinelli, CNR Napoli; S.Riva, CNR Pavia; G. Milanesi, CNR Pavia).
* automation (Luzzana, CNR Milano).
* data bank management (A. Viotti, CNR Milan; C. Saccone, CNR Bari).
* protein analysis (A. Albertini, University of Brescia; A. Fontana CNR Padova).

Because the strategic project is only exploratory there will be no attempt at formal coordination with research other than in Italy (Dulbecco, Toccini-Valentini, Vezzoni), although it is anticipated that any finalized project would be undertaken in the context of, if not as part of, an international project (Vezzoni). Since Italy subscribes to EMBL, it is hoped that access will be given to the DNA sequencer under development there (see 1.2.3).

There is no industrial involvement in the strategic project, not least because there are still very few companies in Italy that would have an interest. But it is hoped that industry would help finance any finalized project (Vezzoni).

4.3 REFERENCES

Burke, D.T., Carle,G.F. and Olson,M.V. Science, 236: 806-812, 1987.

CNR Strategic Project "Mapping and Sequencing of the Human Genome", CNR Rome, 1987.

Dulbecco, R., Salk Institute, personal communication, 9 July, 1987.

Tocchini Valentini,G., Director, CNR Institute for Cell Biology, Rome, personal communication, 7 July, 1987.

Vezzoni, P., Director, CNR Centro Studi di Fisiologia del Lavoro Muscolare, Milano, personnal communication, 13 July 1987.

5 UNITED KINGDOM

5.1 CONTEXT

The UK with its great historical strengths in molecular biology and genetics is, not surprisingly, well-placed intellectually, if not financially, to contribute substantially at least towards mapping the human genome. Indeed, non-governmental funds have enabled mapping work to begin.

Most biomedical research in the UK is funded by government, either through the universities or through the research councils, which receive their budgets through the Department of Education and Science. Financial circumstances are increasingly driving universities to seek industrial support but their mandate is essentially to support curiosity-driven research. The universities have no formal process for coordination of research and only very recently has there been any attempt (by the University Grants Committee) to set priorities.

The research councils provide support both within their own institutions and to university employees for research within loosely defined areas, of which

biotechnology is an area of overlap for the two councils relevant in the present context - the Science and Engineering Research Council and the Medical Research Council. The councils increasingly support precompetitive industrial research in addition to 'basic' research. The Science and Engineering Research Council incorporates a Biotechnology Directorate which is specifically aimed at linking academic research to industrial needs. Both the Department of Trade and Industry and specific companies contribute to projects organized and run by the directorate.

The Medical Research Council by tradition has supported the 'best' proposals it receives rather than directing its funds towards stated research priorities, although it has recently undertaken a directed programme of research into AIDS. The council has a long-standing committment to molecular biology and has the power to set up new units devoted to particular areas of research when a suitable director and sufficient funds are available. A new Secretary (chief executive) of the Medical Research Council will be in place in October 1987 but it is too early to know whether this will herald any increase in the council's support for human genome work.

Both the Department of Health and Social Security and the Department of Trade and Industry have funds available for contract research in areas of

relevance to their needs.

5.2 WORK IN PROGRESS AND PLANNED

5.2.1 Medical Research Council (MRC)

MRC has no current plan to support a coordinated programme of research on human genome mapping/sequence. It already supports a good deal of relevant research, has no defined plan to increase that support and would consider any requests to do so on their merits (MRC).

At a meeting earlier this year, the council endorsed the plan of its employee, Sydney Brenner, to map the human genome largely with private funds (see below) and, at his request, also set up a committee that will consider such questions as the ownership of the cosmids that are produced and how best to provide public access to them. There was general agreement that access should be open to all.

MRC Molecular Genetics Unit

Sydney Brenner, director of this Medical Research Council unit in Cambridge, is adapting the technique he pioneered for mapping the genome of the nematode Caenorhabditis elegans (Coulson) to the human genome.

The nematode genome is the smallest known for any multi-celled creature; it is estimated to be 80 million base pairs compared with 1 billion base pairs for the human genome. Brenner's mapping technique, in essence, starts with a cosmid bank prepared from total genomic DNA. Individual cosmids are characterized in terms of the pattern of restriction fragments they yield on a gel. The patterns are stored in a computer which is programmed to search for similarities. Similarities in pattern serve to identify overlapping cosmids (genome fragments) and thus to allow the assembly of a continuous physical map of the genome. In practice, it is relatively easy to link groups of up to 20 cosmids but increasingly hard to fill in the gaps.

For human genome mapping, Brenner would like to start with the 1000 or so available clones of human DNA, find out to which cosmid each corresponds and use these cosmids as starting points from which to expand segments of the map. Perhaps half of the genome could be mapped in that way by a few people within 5 years. To what extent it would be worth continuing thereafter is open to question, given the increasing difficulties (Brenner).

The terms of reference of Brenner's unit allow the development of techniques that underpin human genome

mapping, for which he has 3 or 4 staff positions, but not the mapping itself. In the first place, Brenner intends to fund that aspect privately. He has access to about US$100,000 from private, largely industrial, sources plus some US$500,000 from prize money. However, he is seeking a more substantial commitment to ensure the continuation of his work after his retirement (Brenner).

Other MRC Units

A number of units and grantholders with interests in specific genetic diseases have a consequent interest in mapping and sequencing particular chromosomal loci. For example, Kay Davies of the MRC Molecular Medicine Unit in Oxford works on Duchenne muscular dystrophy and David Weatherall, director of the unit, has concentrated on thalassemias.

Laboratories with experience and facilities for large-scale (by present standards) sequencing include the Medical Research Council's Institute of Virology in Glasgow and its Laboratory of Molecular Biology in Cambridge but neither has any plans to partake in human genome sequencing. The Cambridge laboratory is in the process of developing gel scanners and sequence software in collaboration with Amersham International.

5.2.2 Imperial Cancer Research Fund

The London laboratory of the Imperial Cancer Research Fund (ICRF) has a considerable interest in mapping human genes. Sir Walter Bodmer, director of the laboratory, has long been involved in mapping the human major histocompatibility complex and his colleagues have particular interests in oncogenes and in the Y chromosome. And recently Bodmer has recruited Hans Lehrach and Anna-Maria Frischauf to pursue a general approach to mapping the human genome.

In theory, their approach is simpler and quicker than Brenner's, but it is untried in practice. Like Brenner, they plan to start with cosmids and to generate a computer-recognized pattern that will allow overlapping cosmids to be identified. In this case, the patterns are to be generated by hybridizing successive synthetic oligonucleotide probes to filters carrying the cosmids (Poustka) Lehrach calculates that 100 probes should suffice. Like Brenner, Lehrach does not foresee completing the map but believes it should not be difficult to link 25 cosmids at a time (Lehrach). Attention will also be given at ICRF to the development of relevant software and instrumentation.

ICRF is a charitable organization that is financed solely (but generously) by donations.

5.2.3 Science and Engineering Research Council (SERC)

SERC supports no research that is directly relevant to human genome mapping/sequencing, although its Protein Engineering Club programme includes work on the prediction of protein structure from sequence (SERC).

5.2.4 Other Work

Considerable relevant work is underway within universities with a variety of sources of funds. For example, Robert Williamson at St Mary's Hospital Medical School in London has a particular interest in cystic fibrosis, and some interest in mapping techniques per se (Vissing) And Andrew Coulson and John Collins at the University of Edinburgh are working on the application of high-performance computers to biosequence analysis (Coulson and Collins).

5.2.5 Departments

None of the relevant government departments has any specific interest in human gene mapping/sequencing.

5.3 COLLABORATION

Talks are planned between the Medical Research Council and the Imperial Cancer Research Fund to explore the common ground between them in human genome research and the possibility of their collaboration

(MRC). There are no plans to collaborate with other countries within or outside Europe at present.

There is no plan to coordinate relevant work going on within the universities but few of the individuals involved are working in a vacuum and many would be glad to participate to some extent in any large project that might emerge.

5.4 REFERENCES

Brenner,S., Director, Medical Research Council Molecular Genetics Unit, Cambridge, personal communication, 26 June and 9 October 1987.

Coulson,A., Sulston,J., Brenner,S. et al. "Toward a physical map of the genome of the nematode Caenorhabditis elegans", Proceedings of the National Academy of Sciences U.S.A. 83: 7821-7825, October 1986.

Coulson,A.F.W. and Collins,J.F. "Biological Sequence Analysis: Advanced Architecture Computing Requirements", paper prepared for Commission of European Communities, XII/604/86, Brussels 1986.

Lehrach,H., Staff, Imperial Cancer Research Fund, London, personal communication, 24 June 1987.

MRC spokesman, Medical Research Council, London, July 1987.

Poustka,A., Pohl,T., Barlow,D.P. et al. "Molecular

Approaches to Mammalian Genetics", Cold Spring Harbor Symposia on Quantitative Biology, Vol LI: 131-139, 1986.

SERC Directory of Research in Biotechnology, Science and Engineering Research Council Biotechnology Directorate, October 1986.

Vissing,H., Grosveld,F., Soloman,E. et al., "Towards construction of a total restriction fragment map of a human chromosome", Nucleic Acids Research 15: 1363-1375, 1987.

6 OTHER COUNTRIES

There are no plans for any independent large-scale projects in human genome sequencing/mapping in any other European country, but in most countries there are individuals carrying out relevant research who would be glad to collaborate in larger projects if they are not already doing so. For example, Charles Weissmann of the University of Zurich has expressed some interest in the exhaustive sequencing of all cDNAs represented in a tissue-specific cDNA library, Michael Steinmetz of Roche in Basel is contributing the cosmids to a large-scale sequencing project of the French Centre d'Etude du Polymorphisme (see 2.2.1), and Peter Pearson in Leiden has considerable interest in the Duchenne muscular dystrophy locus (Weissmann, Pearson, Brenner, Tooze).

REFERENCES

Brenner,S., Director, Medical Research Council Molecular Genetics Unit, Cambridge, personal communication, June 1987.

Pearson,P. State University, Leiden, Netherlands, personal communication, July 1987.

Tooze,J., Secretary, European Molecular Biology

Organization, personal communication, July 1987.

Weissmann,C. University of Zurich, Switzerland, personal communcation, July 1987.

7 DATABASES

7.1 CONTEXT

The growth of databases in Europe, as elsewhere, has been somewhat erratic with the initiative sometimes coming from researchers who have their own need for the information and sometimes from organizations that wish to provide a service. Since European science is not a cohesive entity, some databases have grown nationally whereas others have been catalysed by European needs.

The Commission of the European Communities has recognized the need for "knowledge bases local and global, by which the power of human skills can be amplified. The naturally trans-national character of such infrastructure - the sequence banks, software libraries, cell and microbial collections, seed and tissue banks, for example - again makes collaborative action through Community programmes particularly appropriate" (Narjes).

7.2 EXISTING DATABASES

7.2.1 International

European Commission

The Biotechnology Action Progamme (BAP) of the

European Commission (see 1.1.1) funds several relevant banks and databases. Other funds come through the Commission's Directorate General for Telecommunications, Information Industries and Innovation (Commission...BIOT/5/86) under the influence of its Task Force for Biotechnology Information (Commission...XII/86).

Hybridoma Databank. Founded with support from TFBI and situated at the University of Nice (France), this is the European node of a bank of information on cloned cell lines, in particular hybridomas, which also has nodes in the United States and Japan.

European Human Genetic Mutant Cell Bank. A joint project between the Centre for Applied Microbiology and Research at Porton (UK) and Erasmus University, Rotterdam (Netherlands). Funded by BAP to store and provide mutant cell lines.

Protein Sequence Data Base Funded by BAP at the Max Planck Institute for Biochemistry at Martinsried (FRG), this is the new European node of the long standing US NBRF (Washington) database.

EMBL

There is an unrestricted and open collaboration between the European Molecular Biology Laboratory's

Data Library and the US National Institutes of Health's GenBank in the maintenance of a nucleotide sequence database. The task of entering the data is shared between the two partners and both are trying to encourage the submission of data direct from laboratories and in compatible form. There is considerable concern within the European scientific community regarding the quantity and quality of data in the database, problems which are likely to receive more attention with the expansion of EMBL's biocomputing programme (see 1.2.3).

7.2.2 National

Nationally held databases include:

SWISS-PROT

A protein sequence data bank at the University of Geneva (Switzerland).

PG-TRANS

A protein sequence data bank in the Computer Science Unit of the Pasteur Institute (France).

European Collection of Animal Cell Cultures

Supported by the UK Department of Trade and Industry, this facility is a European counterpart to the American Type Culture Collection of animal cells. It stores and provides cells as well as acting as a recognized patent depository of cell lines. It is

housed in the Centre for Applied Microbiology and Research at Porton (UK).

GENATLAS

Announced in September 1987, this French database will hold information on the locations of all genetic markers, their associations with disease, and bibliographic detail. GENATLAS is based on software created by Professor de Heaulme of the Hopital de la Salpetriere and is run by Professor Jean Frezal of l'Hopital Necker-Enfants Malades.

7.3 PLANS

The need for databases and DNA banks or libraries in the context of hyman genome mapping/sequencing is, of course, well recognized and subject to considerable discussion within Europe but no firm plans have emerged.

7.4 REFERENCES

Commission of the European Communities, Biotech Info in the EEC Work Plan 1986-87-88, BICT/5/86, Brussels,1 October, 1986.

Commission of the European Communities, Bioinformatics: Collaborative European Programme and Strategy, draft XII/86, Brussels, 21 April 1986 (with revised Annex I 30.9.86).

Narjes,H.J. "European cooperation in biotechnology"

Trends in Biotechnology" 5: 118-119, 1987.

7.5 DATABASE ADDRESSES

Hybridoma Data Bank - University of Nice/CODATA, Parc Valrose, 0604 Nice Cedex, France.

European Human Genetic Mutant Cell Bank and European Collection of Animal Cell Cultures - Centre for Applied Microbiology and Research, Porton Down, Salisbury, Wilts SP4 0JG, UK.

Protein Sequence Data Bank - Max-Planck-Institut fur Biochemie, D-8033 Martinsried, FRG.

EMBL Data Library - EMBL, Postfach 10.2209, D-6900 Heidelberg 1, FRG.

SWISS-PROT - Department de Biochemie Medicale, Universite de Geneve, 1211 Geneve, Switzerland.

PG-TRANS - Computer Science Unit, Institut Pasteur, 28 rue de Dr Roux, 75724 Paris, Cedex 15, France.

8 REFERENCES

Beck,S. and Pohl,F.M., "DNA sequencing with direct blotting electrophoresis" EMBO J. 3: 2905-2909,1984.

Brenner,S., Director, Medical Research Council Molecular Genetics Unit, Cambridge, personal communication, 26 June and 9 October 1987.

Burke, D.T., Carle,G.F. and Olson,M.V. Science, 236: 806-812, 1987.

Buschbeck, F., Bundes Ministerium fur Forschungs und Technologie, Bonn, personal communication, July 1987.

Chambon,P., Faculty of Medicine, University of Strasbourg, personal communication, July 1987.

Claverie,J.-M., Institut Pasteur, Paris, personal communication, July 1987.

Cohen, D., Staff, Centre d'Etude du Polymorphisme Humain, Paris, France, personal communication, July and October 1987.

Commission of the European Communities, Biotech Info in the EEC Work Plan 1986-87-88, BIOT/5/86, Brussels, 1 October, 1986.

Commission of the European Communities, Biosequences: Perspectives and User Services in Europe, XII.604.86, Brussels, 1986.

Commission of the European Communities, Bio-informatics: Collaborative European Programmes and

Strategy, draft XII/86, Brussels, 21 April 1986 (with revised Annex I-30.9.86).

Commission of the European Communities, List of Research Contracts in the Biotechnology Action Programme, XII/87-BN/hl, Brussels, 7 June 1987.

Coulson,A., Sulston,J., Brenner,S. et al. "Toward a physical map of the genome of the nematode Caenorhabditis elegans", Proceedings of the National Academy of Sciences U.S.A. 83: 7821-7825, October 1986.

Coulson,A.F.W. and Collins,J.F. "Biological Sequence Analysis: Advanced Architecture Computing Requirements", paper prepared for Commission of European Communities, XII/604/86, Brussels 1986.

CNR Strategic Project "Mapping and Sequencing of the Human Genome", CNR Rome, 1987.

Dausset,J., Cann,H. and Cohen,D. "Centre d'Etude du Polymorphisme Humain (CEPH): Collaborative Mapping of the Human Genome" background paper CEPH, College de France, Paris , undated.

Deutscher Bundestag "Chancen und Risiken der Gentechnologie",d.Bericht d. Enquete-Kommission d. 10 Deutscher Bundestages, Deutsches Bundestag, Referat Offentlichkeitsarbeit Bonn, 1987.

Dulbecco, R., Salk Institute, personal communication, 9 July, 1987.

EMBL Annual Report 1985

EMBL Annual Report 1986

EMBL, The 1987-1990 Scientific Programme of the EMBL, EMBL/86/3 rev. 1 E, Heidelberg, 1 September 1986.

Flensholt, J., Jacobsen, H.K. & Jakobsen, L.K. "Proposal for a European Action Plan in Bio-Informatics" draft report prepared for the Commission of the European Communities BICEPS, Brussels, December 1986.

Goffeaut, A., DG XII, Commission of the European Communities, Brussels, personal communication, July 1987.

Hofschneider,P., Director, Max Planck Institut fur Biochemie, Martinsreid, personal communication, 14 July 1987.

Lehrach,H., Staff, Imperial Cancer Research Fund, London, personal communication, 24 June 1987.

Mandel,J.-L., Faculty of Medicine, University of Medicine, personal communication, July 1987.

Martin,W.J., Warmington,J.R., Galinski,3.R. et al. "Automation of DNA Sequencing" Bio/technology 3:911-915, 1985.

Martin, W. J. & Davies, R.W., "Automated DNA Sequencing: Progress and Prospects" Bio/technology 4: 890-895, October 1986.

MRC spokesman, Medical Research Council, London, July 1987.

Narjes,H.J. "European cooperation in biotechnology" Trends in Biotechnology" 5: 118-119, 1987.

Nieuwenhuis, B., DGXII Commission of the European Communities, Brussels, personal communication, 7 July and 8 October 1987.

Oliver, S., Director, Manchester Biotechnology Centre, University of Manchester Institute of Science and Technology, Manchester, UK, personal communication, July 1987.

Pearson,P. State University, Leiden, Netherlands, personal communication, July 1987.

Philipson,L., Director-General, European Molecular Biology Laboratory, Heidelberg, FRG, personal communication, July 1987.

Poustka,A., Pohl,T., Barlow,D.P. et al. "Molecular Approaches to Mammalian Genetics", Cold Spring Harbor Symposia on Quantitative Biology, Vol LI: 131-139, 1986.

Rechaussat,D. INSERM, personal communication, October 1987.

SERC Directory of Research in Biotechnology, Science and Engineering Research Council Biotechnology Directorate, October 1986.

Thomas,D., Technical University of Compiegne, personal communication, July 1987.

Tocchini Valentini,G., Director, CNR Institute for Cell Biology, Rome, personal communication, 7 July, 1987.

Tooze,J., Secretary, European Molecular Biology Organization, Heidelberg, FRG, personal communications, June and July 1987.

Vezzoni, P., Director, CNR Centro Studi di Fisiologia del Lavoro Muscolare, Milano, personnal communication, 13 July 1987.

Vissing,H., Grosveld,F., Soloman,E. et al., "Towards construction of a total restriction fragment map of a human chromosome", Nucleic Acids Research 15: 1363-1375, 1987.

Weissmann,C. University of Zurich, Switzerland, personal communcation, July 1987.

Zachau,H., University of Munich, personal communication, 26 June 1987.

UNIVERSITY OF OXFORD

PROFESSOR SIR DAVID WEATHERALL FRS
Telephone OXFORD 81-7632
OXFORD 60201

NUFFIELD DEPARTMENT OF
CLINICAL MEDICINE
JOHN RADCLIFFE HOSPITAL
HEADINGTON
OXFORD OX3 9DU

17th July, 1987

Robert Cook-Deegan,
Senior Analyst and Project Director,
Congress of the United States,
Office of Technology Assessment,
Washington
DC 20510-8025

Dear Mr. Cook-Deegan,

Please find enclosed my first effort for your report. My secretary happens to be visiting the USA this week and this will be posted in the USA so it should meet your deadline.

Since I was still slightly unsure what you required I hope that this initial effort will be useful to you. I look forward to hearing from you.

Yours sincerely,

D.J. Weatherall.

APPLICATION OF HUMAN GENE MAPPING FOR THE GLOBAL CONTROL OF GENETIC DISEASE

D.J. Weatherall, M.D., F.R.C.P., F.R.S.

MRC Molecular Haematology Unit,

Nuffield Department of Clinical Medicine,

University of Oxford,

John Radcliffe Hospital,

Oxford, OX3 9DU

INTRODUCTION

One of the major reasons for mapping and ultimately sequencing the human genome is the control of genetic disease in man. In assessing the importance of this project in this context it is important to consider how information derived from the knowledge of the sequence of the human genome might be applied globally for the management of genetic diseases. Because the first disorders of this type to be characterized at the molecular level, the hemoglobinopathies, are the commonest genetic diseases in the world population, and because programs for their control have been initiated in parts of the developing world, it is already possible to define some of the problems that will arise in applying knowledge gained from a genome sequencing program to the improvement of the health of populations in which genetic diseases are particularly common.

Here I shall outline what is known about the world distribution of single gene disorders and about the genetic component of important polygenic conditions in the developed and developing countries, and consider briefly the organisational difficulties that may be encountered in applying this new information to programs for the prevention, and ultimately the treatment, of genetic disease.

THE GLOBAL PROBLEM OF GENETIC DISEASE

The main groups of genetic disorders are summarized in Table 1. First, there are single gene defects that may follow a dominant, recessive, or sex-linked pattern of inheritance. Second, there are the chromosomal abnormalities, that can be subdivided into those in which there is variation in chromosomal number and those in which there is a chromosomal rearrangement. The third group is comprised of congenital malformations, at least some of which seem to have a strong

Table 1 The estimated genetic load in developed north European populations (from Weatherall, 1985 which contains the original sources for these data).

Type of genetic disease	Frequency/ 1000 population
Single gene	
Dominant	1.8 - 9.5
Recessive	2.2 - 2.5
X-linked	0.5 - 2.0
Chromosome abnormalities	6.8
Common disorders with a significant genetic component	(7 - 10)
Congenital malformations	(19 - 22)
Total (approximate)	(37.3 - 52.8)

genetic component; it is clear, however, that environmental factors play a major role in the pathogenesis of many disorders which come under this general heading. Finally, there is a group of common diseases in which genetic factors are thought to play a role, although for which the relative contribution of genetic and environmental mechanisms is not yet clear.

The Total Burden of Genetic Disease

A breakdown of the overall frequency of genetic disease is shown in Table 1. These data represent only a crude estimate of the burden of genetic disease in north European populations; they suggest that approximately 1-2% of all births are associated with some form of genetic disability or developmental disorders with a strong genetic component.

Another approach to assessing the importance of these disorders is to analyse the frequency of genetic disease and congenital malformation in patients in paediatric hospitals; an excellent example of a study of this type was reported from Montreal by Scriver et al (1973). In an analysis of a sample of 12801 admissions between 1969 and 1970, 'genetic' admissions accounted for 11.1%, and congenital malformations for 18% of the total. The overall estimate was that about a third of all admissions were for diseases with a genetic component. Studies of this type provide clear evidence that genetic disease puts a serious burden on the health services in developed countries. Subsequently several studies have confirmed these findings, and in addition, have shown that genetic disease is a major cause of death under the age of 15 years (Weatherall, 1985a).

Racial Variation in The Prevalence of Genetic Disease

The data shown in Table 1 describes the approximate burden of genetic disease for individuals of north European origin. However, the relative frequency of these conditions varies considerably among populations. Some indication of the variation in the frequency of monogenic disorders between different racial groups is shown in Table 2. Well documented examples include the high frequency of cystic fibrosis in north Europeans as compared with Afro-Americans or orientals, and the extremely high frequency of Tay Sachs disease among Ashkenazi Jews.

Of the conditions listed in Table 2, the genetic disorders of the red cell, thalassemia and the common structural hemoglobin variants, are by far the most common. Information obtained over the last few years suggests that these conditions are probably the commonest single gene disorders in he world population (Weatherall and Clegg, 1981; WHO, 1982). Since these diseases are so well understood at the phenotypic and molecular levels, and because so much is known about their world distribution, they provide an important model system for analysing the problems of applying DNA technology globally for the control of genetic disease.

The Hemoglobin Disorders and Other Red Cell Polymorphisms As A World Health Problem

Globally, the important hemoglobin disorders are the α and β thalassemias and the β globin structural variants, hemoglobins S, C, and E. The homozygous states for some forms of α thalassemia, β thalassemia and hemoglobin S (sickle cell anemia) produce serious clinical disability. In addition, the compound heterozygous states for β thalassemia and hemoglobins S, C, or E are also usually

Table 2 The racial differences in frequency of genetic disease (from Weatherall, 1985 which contains the sources of these data).

Disease	Race	Frequency/ 1000 births
Porphyria variegata	South African (White)	3.0
	Caucasians (General)	0.01
Huntington's chorea	Tasmania	0.17
	Japan	0.03
Adrenogenital syndrome	Yupik Eskimos	2.0
	N. Americans	0.025
Cystic fibrosis	N. Europeans	0.4 - 0.5
	Afro-Americans. Orientals	0.01
Tay-Sachs disease	Ashkenazi Jews	0.17 - 0.4
	Shephardi Jews. Gentiles	0.001 - 0.003
Thalassemia	Mediterraneans. Orientals	10 - 20
Sickle-cell anemia	Africans	10 - 20

associated with a severe clinical phenotype.

An account of the clinical manifestations and molecular genetics of these disorders for the non-specialist is given in the monograph of Weatherall (1985a) and an up-to-date and detailed review of the field will be found in the monograph of Bunn and Forget (1986). Essentially, all these disorders involve the globin chains of adult hemoglobin. This molecule has two α chains and two β chains ($\alpha_2\beta_2$). The structural hemoglobin variants are the result of single amino acid substitutions in these chains while the α and β thalassemias result from a variety of molecular defects that cause defective synthesis of either the α or β chains.

Although our knowledge of the prevalence and distribution of these conditions is still scanty enough information has been obtained to suggest that they will constitute an increasingly serious world health problem over the next half century (Livingstone, 1967, 1985; Weatherall and Clegg, 1981; WHO, 1982, 1983; Bunn and Forget, 1986).

Currently, about 50% of the world's population consists of groups with about 5% of carriers of potentially pathological hemoglobin variants; about 190 million people, 4% of the world population, carry a gene for a major hemoglobinopathy. Although more than half of these are thalassemia genes it is likely that globally there are more births of infants with sickle cell disease than with the thalassemia syndromes because of the very high density of sickle cell carriers in Africa and parts of India. A very conservative estimate (WHO, 1982; 1983) suggests that there are about 250-300,000 births a year of children with the β chain hemoglobinopathies.

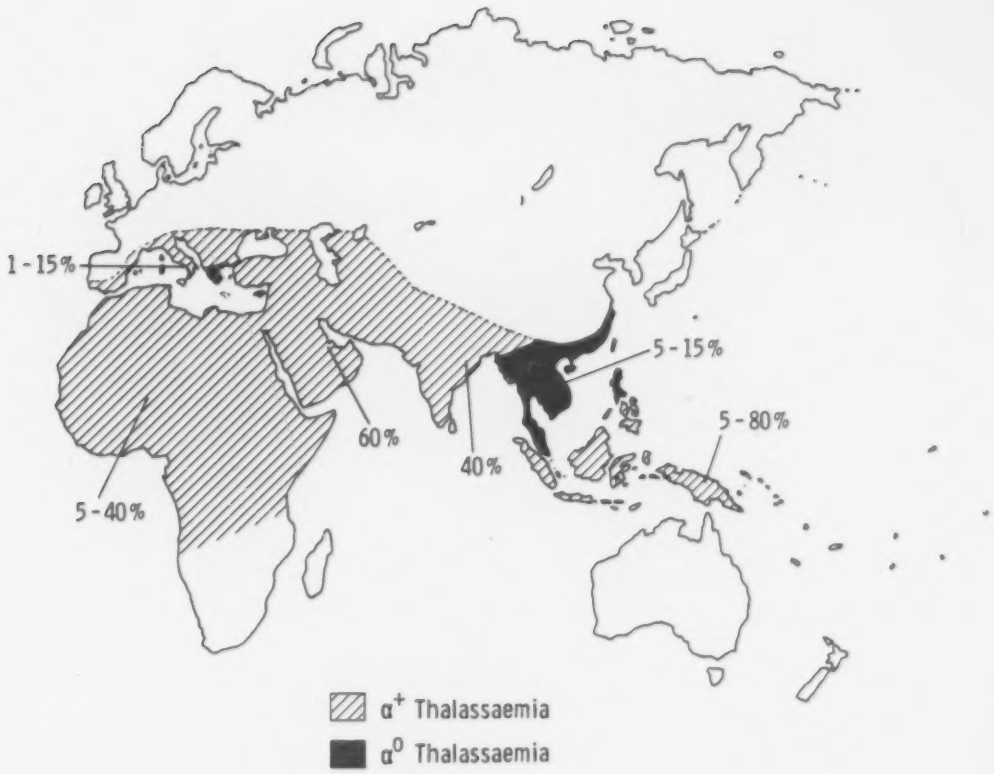
Recent migrations have introduced these conditions into populations in which they were not originally endemic. For example,

in northwest Europe between 2 and 9% of most populations now belong to ethnic groups at high risk for these disorders. In some countries, population movements and admixture will reduce the number of homozygous births by dispersing localized groups with high heterozygote frequencies. On the other hand, in southeast Asia, where the hemoglobin E gene predominates in one part of the population and β thalassemia in another, population mixing may well have the opposite effect. In Thailand, for example, population admixture could increase the potential birth rate of infants with homozygous β thalassemia or hemoglobin E thalassemia by as much as 30% (WHO, 1983).

It is estimated that the world population may stabilize around the year 2100 at more than twice its present size (United Nations Secretariat, 1982). The fastest growing populations are those with the highest incidence of carriers for important hemoglobin variants. Ultimately up to 70% of the world population will be in the high risk groups for the hemoglobinopathies, and up to 6% may be heterozygotes.

The Sickling Disorders. The sickle cell gene is most concentrated in west Africa although its distribution spreads across central Africa and it occurs at lower frequencies in some non-African Mediterranean populations. It occurs patchily throughout the Middle East where it reaches particular high frequencies in eastern Saudi Arabia. Recent studies (Kar et al, 1986) indicate that it is extremely common in parts of central India, but apart from some Indian populations in north Malaya, it has not been observed in southeast Asia. In parts of west central Africa, the gene frequency for hemoglobin S exceeds 0.15; in other words heterozygotes make up 25% of the newborn population. Similar or higher frequencies are found in parts of eastern Saudi Arabia.

Figure 1 The world distribution of the α thalassemias.

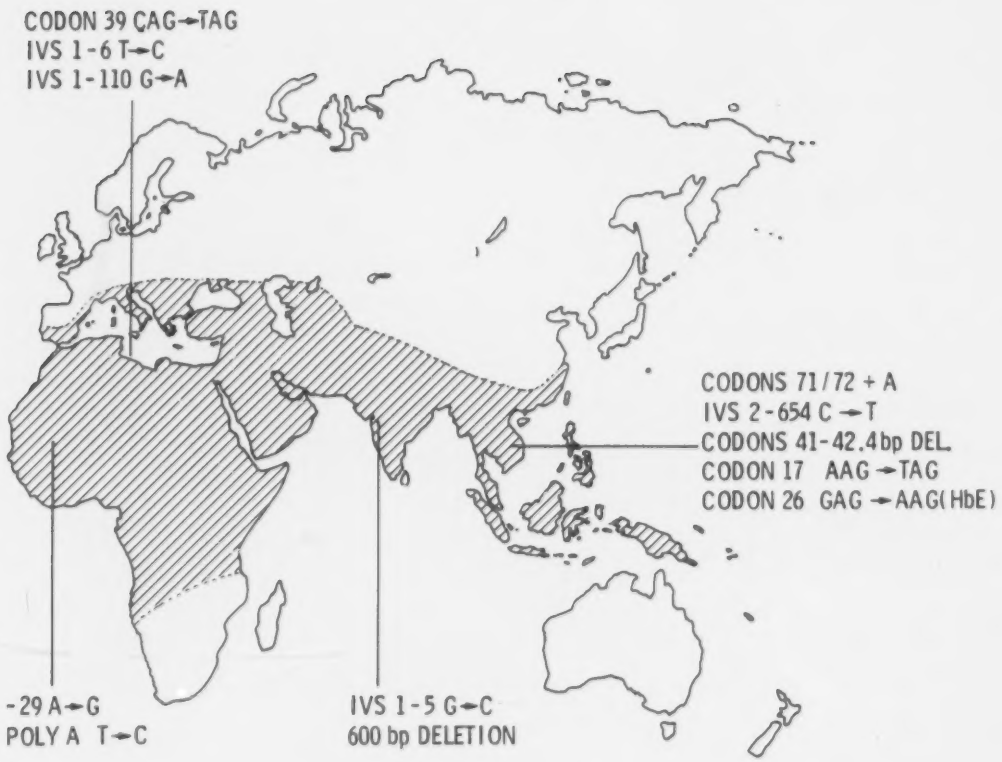


Hemoglobin E. The gene frequency for hemoglobin E in Thailand and Burma is 0.05-0.10 although higher values have been recorded in parts of eastern Thailand near the Vietnamese border, to so-called hemoglobin E triangle. At a minimum estimate there are about 30 million heterozygotes for hemoglobin E, and about 1 million homozygotes in southeast Asia.

The Thalassemias. The α thalassemias are classified into the α^{0} thalassemias, in which both α globin genes on chromosome 16 are deleted, and the α^{+} thalassemias in which one of the two pairs of α globin genes on chromosome 16 are either deleted or inactivated by point mutations. The homozygous state for α^{0} thalassemia causes fetal death either late in pregnancy or immediately after birth and is associated with a very high incidence of maternal complications in pregnancy. The homozygous state for α^{+} thalassemia causes a mild asymptomatic hypochromic anemia. The compound heterozygous state for α^{0} and α^{+} thalassemia produces a moderately severe hemolytic anemia called hemoglobin H disease.

The world distribution of the α thalassemias is shown in Figure 1. α^{0} thalassemia is restricted to southeast Asia and the Mediterranean region while the α^{+} thalassemias are by far the commonest single gene disorders in man; their frequency ranges from about 30% in various African and Middle Eastern populations up to 80-90% in the coastal regions of Papua New Guinea, where the gene seems to be moving towards fixation. Although α^{+} thalassemia does not cause severe anemia in the homozygous state, its very high prevalence causes large populations to have hypochromic red cells. For example, in some of the southwest Pacific islands 40-50% of individuals have hypochromic blood films nearly all due to α thalassemia rather than iron deficiency (Bowden et al, 1985). At a time when there is a major

Figure 2 The world distribution of the β thalassemias. The mutations that are shown all involve the β globin genes. IVS indicates intervening sequence. The numbers indicate the position of the particular mutation. DEL indicates deletion.



effort to fortify diets with iron it is essential that the frequency of these genetic causes of hypochromic anemia are ascertained.

The β thalassemias are very heterogeneous at the molecular level; over 40 different underlying mutations have been found (reviewed by Orkin and Kazazian, 1984 and Weatherall et al, 1988). Adequate gene frequency data for β thalassemia are only available for a small number of populations (Weatherall and Clegg, 1981; Livingstone, 1985). The disease is particularly common throughout the Mediterranean region, parts of the Middle East, and India and Burma. It also occurs throughout southeast Asia in a line starting in southern China, stretching through Thailand, Cambodia and Laos, and down the Malay peninsula into some of the island populations; it is distributed sporadically in Melanesia.

Recent studies suggest that the different mutations that are responsible for β thalassemia are not evenly distributed throughout the world population. Rather, each high incidence population seems to have two or three common mutations together with a variable number of rare forms. It follows that many apparent homozygotes are in fact compound heterozygotes for two different molecular lesions. The world distribution of β thalassemia and some of the common mutations in different population groups are summarized in Figure 2.

Glucose-6-Phosphate Dehydrogenase Deficiency. The world distribution of glucose-6-phosphate dehydrogenase deficiency is very similar to that of the hemoglobinopathies (Luzzatto and Testa, 1978). Although this X-linked condition does not cause chronic ill health it is important because it leads to sensitivity to a variety of oxidant agents, notably many of the commonly used anti-malarial drugs. With the rapid resurgence of malaria in many parts of the world, and the increasing development of multiple drug resistance, glucose- 6-

phosphate dehydrogenase deficiency will assume a greater public health importance. Furthermore, in some populations it is an important cause of neonatal jaundice and kernicterus, with subsequent brain damage.

The Health Burden And Economic Importance Of The Common Genetic Disoders Of The Blood For The Developing World

Because of the high neonatal and infant mortality rates due to infection and malnutrition in many third world countries genetic disease is often not recognised as a major clinical problem. However, as social circumstances and medical services improve, these conditions start to appear with increasing frequency in affected populations. The problem that a very common genetic disease can impose for a community is well illustrated by some cost analyses that were made for thalassemia in Cyprus (WHO, 1982). This condition was almost unknown in Cyprus until the early 1950s. Following a WHO malaria eradication program just after the Second World War, together with the improved socio-economic conditions that occured at this time, the death rates in the neonatal period and in early childhood declined. It then became apparent that β thalassemia was a major problem on the island. It has been estimated that if all children with severe transfusion-dependent thalassemia are to be kept alive and treated by regular blood transfusion and appropriate drug therapy to remove excess iron derived from transfused blood, in 20 years time half the entire health budget of the island will be directed towards treating this one disease (Weatherall and Clegg, 1981; WHO, 1982). Similar problems have arisen in Thailand where it is now estimated that there are nearly half a million seriously affected children in a largely rural community that has a relatively well developed national health program. Based on the current Thai data there could well be over a million seriously affected children in southeast Asia by the end of

the century.

Monogenic Disorders That Occur At High Frequencies In Other Populations

Currently, there is no evidence to suggest that there are other single gene disorders in other populations that approach the frequencies of the genetic disorders of the red cell. Some of the discrepancies between the frequency of these conditions in different racial groups is summarised in Table 2. It appears that cystic fibrosis, with a carrier rate of approximately 5%, is by far the commonest single gene disorder in individuals of north European origin. But as mentioned earlier many serious genetic diseases may still go unrecognised in developing countries with a high neonatal and infant mortality rate. And undoubtedly there will be island isolates with a particular high incidence of individual single gene disorders due to founder effect and gene drift. An interesting example of this type of phenomenon is the very high frequency of the gene for oval red cells among Melanesian populations (Kidson et al, 1981); this condition, like the thalassemias, seems to provide protection against malaria. Othe examples of very high frequencies of genetic diseases in Pacific Island populations are described by Bowden et al (1985). Current information about the world distribution of other single gene disorders is given by Vogel and Motulsky (1982).

Polygenic Disease and Developmental Abnormalities

The major causes of mortality in the developed western societies are degenerative arterial disease, cancer, diabetes, autoimmune disease, and the major psychiatric disorders. There is increasing evidence that many of these conditions will become more common in the developing world as social conditions improve (King, 1987). Compared

with the monogenic disorders outlined in the previous sections the value of recombinant DNA technology for the study and control of these conditions is less obvious. With a few rare exceptions none of them can be traced through families in the way suggesting the action of a single gene. Rather, they appear to be polygenic and to reflect the involvement of a number of environmental factors working on a complex pattern of genetic susceptibility (Weatherall, 1985a; Bodmer, 1986).

There are two reasons why it is important to explore the genetic component of these common diseases. First, by defining better the individual genes that constitute these polygenic systems, it should be possible to learn more about their underlying pathophysiology. This is absolutely essential if better forms of treatment are to be developed. Second, there is the strong possibility that if it were possible to define subsets of populations at particularly high risk for developing these conditions it might be possible to concentrate and hence to make more efficient use of preventative programs that entail the removal of environmental risk factors.

Cardiovascular disease presents the major challenge and is a valuable model of a polygenic system that can be explored with DNA technology (Brown and Goldstein, 1986; Deeb et al, 1986). Ischaemic heart disease and cerebrovascular disease result from atheromatous changes in blood vessels, that is the presence of fatty plaques along their endothelial lining. The cause of atheroma is not known. A current working hypothesis is that atheromatous lesions result from a sequence of events starting with endothelial injury and platelet aggregation, followed by release of growth factors from the platelets that cause proliferation of muscle cells in the arterial wall and a local increase in vessel wall permeability to plasma lipoproteins. These complex interactions lead eventually to plaque formation,

fibrosis and ulceration. Although factors such as vascular spasm and platelet aggregation may play a role, it seems likely that the major cause of vascular occlusion and death in patients with widespread atheroma is thrombosis occurring on top an ulcerated atheromatous plaque. Thus in analysng the pathogenesis of cardiovascular death it is insufficient to simply consider the pathogenesis of atheroma; many other factors, including both platelets and the entire blood clotting system and both local and circulating agents that are designed to inhibit unwanted clotting, may also be involved.

The major environmental factors involved in the pathogenesis of vascular disease include dietary fat and possibly carbohydrate, obesity, hypertension, smoking, and many others. There is increasing evidence that a major risk factor is the level of plasma LDL cholesterol (see Brown and Goldstein, 1986). There is no doubt that genetic factors play an important role in setting the level of plasma cholesterol; the most convincing evidence that this is the case is the high incidence of coronary artery disease in individuals with monogenic hypercholesterolemia, a condition that results from a variety of different mutations involving the LDL receptor gene (Brown and Goldstein, 1986). But there is abundant evidence that many other genes must be involved in cholesterol matabolism including those that control the different apolipoproteins involved in its transport, the lipoprotein lipases involved in the breakdown of fatty particles, and the complex enzyme systems controlling its synthesis and degradation. The genes for many of these proteins have now been isolated and many of them show a number of restriction fragment length polymorphisms (RFLPs). So even if we restricted our study of vascular disease to the regulation of cholesterol we already have an abundance of so-called candidate genes, that is genes that could be involved in susceptibility to atherosclerosis. And, as already mentioned, there

are many others.

Having decided on a number of candidate genes and defined related restriction fragment length polymorhisms what is the next step? There are two main approaches (see Bodmer, 1986). The first is to compare the distribution of particular polymorphisms in groups of patients with premature vascular disease with non-affected controls. The second, and more effective, is to examine families with premature vascular disease and look for segregation of particular polymorphisms with disease; this can be done by conventional pedigree analysis or by studying segregation in sibling pairs. Of course there are many pitfalls, particularly genotype assignment, and many mistakes will be made. However, if the association is strong enough it should show up on appropriate statistical analysis. Once having determined that a polymorphism has a genuine association with vascular disease the linked candidate gene and its product can then be analysed in detail.

The same general approach can be used to study genetic factors involved in other common polygenic diseases (Bodmer, 1986; Weatherall, 1987). For example in the molecular analysis of malignant transformation the identification of cellular oncogenes (Bishop, 1987) and the definition of mutations and gene rearrangements with the potential for the modification of oncogene activity provides enormous opportunities for studying the pathogenesis of different human cancers. And even in the most intractable of human diseases, the major affective psychoses and schizophrenia, there has already been some progress in defining candidate genes by the study of rare families in which these disorders appear to follow a monogenic pattern of inheritance (Anderton, 1987).

A similar approach can be used to study the defects of human development and premature ageing. For example, the identification of

DNA markers spread over the human genome provides the opportunity for identifying submicroscopic chromosomal deletions that may be associated with maldevelopment (Weatherall et al, 1981). Standard RFLP analysis of kindreds with more than one individual affected with dysmorphic features should make it possible to define genes that are involved in development (Moore et al, 1987). The recent identification of the gene for the amyloid protein that is laid down in the brains of children with Down syndrome and adults with Alzheimer's disease, and its assignment to chromosome 21, provides an excellent example of the importance of studying candidate genes for polygenic disorders of completely unknown aetiology (Kang et al, 1987).

The Value Of The Analysis Of Polygenic Disease For the Developing World

Most of the examples cited in the previous section have dealt with conditions that are major public health problems in the developed world. However, there is no doubt that with rising standards of living, these disorders will become a major public health problem throughout the world populations (see King, 1987). In addition, in trying to understand some of the important global health problems, particularly infectious disease and parasitic disorders, it will be important to attempt to define genetic factors for susceptibility or resistance to infection. With the exception of the hemoglobin disorders very little is known about genetic susceptibility or resistance to the major bacterial, viral and parasitic illnesses, a problem that is highlighted by our total ignorance about the reasons for the curious discrepancies in the population distribution of AIDS. Indeed, this is a very neglected area of population genetics, largely because until recently the technology for asking questions of this

type did not exist. However, with the availability of increasing numbers of DNA polymorphisms, and by defining a number of candidate genes including those of the immune system, it should now be possible to try to dissect some of the genetic factors which are involved in modifying patterns of infection in populations that undoubtedly exist. The recent study clearly relating the high prevalence of α thalassemia in the Pacific island populations to resistance against P. falciparum malaria is a good example of the enormous power of DNA technology in this branch of population genetics (Flint et al, 1986).

THE VALUE OF GENOME MAPPING AND SEQUENCING FOR THE CONTROL OF MONOGENIC AND POLYGENIC DISEASE

There is already abundant evidence about the value of the practical applications of gene mapping. For example it has allowed us to define many restriction fragment length polymorphism linkage markers close to defined genes that are involved in inherited disorders of known cause. It has also led to the identification of linked markers for a variety of diseases of unknown aetiology including Huntington's chorea, adult polycystic disease of the kidney, the muscular dystrophies, cystic fibrosis, neufibromatosis, and mseveral other single gene disorders. It has made it possible to identify the roles of a number of oncogenes and their specific associations between chromosomal translocations and leukemias or lymphomas. It has demonstrated the importance of homozygosity for certain recessive cancer genes such as those causing retinoblastoma and Wilm's tumor. And, finally, it has also provided clear evidence of heterogeneity of several genetic diseases.

Thus, using restriction fragment length polymorphisms and, in particular, probes for hypervariable regions (Jeffreys et al, 1985) it

is now possible to obtain linkages for well defined single gene disorders provided that the carrier states can be identified and appropriate families are available. This approach is feasible for defining linkages of intermediate distances involving recombination fractions between 5 and 25%. But a 10% recombination fraction corresponds to about 10^7 base pairs or about 1/300th of the genome, which may represent for 20 to 30 gene clusters. Currently, using pulse-field gel electrophoresis, which can separate DNA fragments of 200 to 300 kb or more, it is not possible to cover regions of genome of 10^6 base pairs or more. One million base pairs are covered by 2; non-overlapping cosmids. And even though new approaches are being developed for semi-automatic alignment of randomly chosen cosmids by matching the sizes of simple restriction digest patterns, there is still a gap in the size of the DNA pieces that can be cloned and those that can be separated by physical means. The smallest chromosome fragment that can be identified is about 10^7 base pairs long but the largest cosmid is only 40,000 base pairs; a procedure is needed for cloning pieces midway between these two limits. These considerations mean that finding a particular gene by linkage analysis is extremely tedious and a great deal of luck is involved!

When we come to the analysis of common polygenic disorders using candidate genes the condition is even more complicated. As mentioned in the previous section this entails either individual family studies or the identification of linkage disequilibrium by population studies. As pointed out by Bodmer (1986), unless the DNA-detected variation is sufficiently close i.e. has a recombination fraction of about 5% or less, it will probably not show as a population association. On the other hand if a putative predisposition marker has been found with a 10% recombination fraction it might then be possible to saturate the relevant region with further polymorphic markers and look for one that

shows a clear population association with the trait, rather than by increasing the numbers of families analysed to search for closer linkage. This, in the long run, may be the most efficient way of finding closely linked markers because family data are very inefficient at distinguishing between small recombination fractions e.g. 0.5% compared with 5%; the rate of decline of linkage disequilibrium with a generation number n, $(1-r)^n$, magnifies the effect of a small recombination fraction (r) in population association data. Only the availability of an unlimited range of polymorphic DNA markers would make it possible to pursue these approaches using the methods of classical genetics.

It is clear, therefore, that for the identification of gene loci for single gene disorders, and to facilitate the definition of the different genes involved in the polygenic systems that underlie many common diseases or resistance to them, a complete map of the human genome would be invaluable. Of course, for the control of genetic disease this information has the potential for other important advances. Once we have a complete sequence of the important human gene clusters it will be much easier to identify critical regulatory regions, the enhancer elements for example. Until we have this information it may be difficult to move towards specific gene therapy as a definitive approach for the treatment of common genetic disease. Since the project would provide a stimulus for the development of both appropriate automated technology and the building up of large data bases, it would be relatively simple to apply this information rapidly towards the practical goals of the control and management of genetic disease.

PRACTICAL APPROACHES TO THE CONTROL OF GENETIC DISEASE IN MAN

The establishment of a complete map of the human genome, together with the collation of information about the sequence of large regions of importance, should allow us to obtain information about the chromosomal location and molecular pathology of most of the important single gene disorders in man and provide us with considerable insight into the genetic component of polygenic systems. Given this information how would we apply it for the control of human genetic disease?

Apart from mutagen control for the primary prevention of genetic defects, there are a number of approaches to the secondary control of common genetic diseases. First, it is possible to screen populations for genetic diseases; once carriers are detected they can be given appropriate marital advice in the hope that they will avoid mating with other carriers. This type of secondary prevention of genetic disease would only be cost-effective in populations in which particular genetic disoders occur at a high frequency. Although there is very little experience of the population control of single gene disorders using this approach, such that there is suggests that it will not be very effective (Stamatoyannopoulos, 1973; Weatherall, 1985a).

Another approach to the secondary prevention of single gene disorders is to carry out maternal screening in early pregnancy; when the carrier status for a serious genetic disease is identified the husband can be screened and then the couple counselled about the risk of having an affected child. If they so wish prenatal diagnosis can be carried out, either in the first trimester by chorion villus sampling or later in pregnancy by amniocentesis or fetal blood sampling. If the fetus is affected the couple can opt for termination

of the pregnancy.

Finally, neonatal screening programs can be developed, but they are only of value for conditons such as phenylketonuria where treatment of a genetic disease is possible once an affected infant is discovered. Unfortunately the treatment of most genetic diseases is very unsatisfactory and therefore this method of screening has a very limited place in the control of genetic disease.

Until specific gene therapy is developed it seems likely that the major approach to the control of genetic disease globally will lie in the detection of the disorder in mothers early in pregnancy and the application of prenatal diagnosis. Since first trimester prenatal diagnosis by chorion villus sampling and DNA analysis is already of proven value I shall consider its application in some detail.

Prerequisites for A First Trimester Prenatal Diagnosis Program Using Fetal DNA

The general approach to developing a prenatal diagnosis program for a specific genetic disease using fetal DNA analysis has been outlined in detail recently (Weatherall et al, 1987). First, either the specific gene locus involved in the disease must be identified or, if this is not possible, restriction fragment length polymorphisms close enough to the locus to make prenatal diagnosis possible must be defined. The next step is to determine the genetic and molecular heterogeneity of the disorder. In other words, it is essential to examine sufficient families to make absolutely sure that there is only one locus involved as the basis for a particular genetic disease. In addition it is important to determine the molecular heterogeneity and how this relates to the clinical phenotype. This is essential for genetic counselling purposes. It may turn out, in addition, that the

phenotype can be modified by the action of one or more genetic determinants either in the same gene cluster or in a distant cluster, as is the case for some of the hemoglobin disorders. Having determined the molecular pathology and heterogeneity it should then be possible to construct oligonucleotide probes for direct analysis of the mutation, both for carrier detection and prenatal diagnosis. And the final step, which may be particularly important if the prevention program is to be cost-effective, is to attempt to work out a simple biochemical screening test for the carrier status for the particular genetic ddisorder.

Since to date all these criteria have only been met for the human hemoglobin disorders it might be useful to determine how a prenatal diagnosis program for these conditions can be set up, both in the developed world and in the developing countries where these conditions are particularly common.

A Comprehensive Prenatal Diagnosis Program For The Hemoglobin Disorders.

Since these disorders are the commonest single gene conditions in man, and because they reach their highest frequencies in parts of the world that are still relatively under-developed, it is clear that the simplest and most economical approaches must be taken.

The first requirement for a program for prenatal diagnosis of the hemoglobinopathies is an adequate diagnostic service for accurately defining heterozygotes. This is relatively easy; the methods and technical problems are reviewed in detail elsewhere (Weatherall, 1985b) and the techniques can be applied in simple rural community hospitals or health centres. The second preliminary stage involves a major educational program to determine whether the population wants

prenatal diagnosis and to ensure that they understand the significance of genetic disease for their society.

Once having embarked on a major public health measure of this type the next problem is to define the molecular pathology of the common hemoglobin disorders in the population. From a practical point of view these will be the α and β thalassemias and either the sickle cell or hemoglobin E genes. Since both α and β thalassemia are heterogeneous at the molecular level, and because individual mutations seem to be confined to particular populations (Figures 1 and 2), the most economical approach to working out the molecular pathology of the thalassemias in populations in which information about the common mutations is already available (see Figure 2) is by the use of oligonucleotide probes constructed to identify individual base changes. This is a fairly sophisticated operation and may well require the services of one or two central laboratories, possibly under the auspices of the WHO or a similar body. If it appears that there are common mutations in a particular population that are not identified by oligonucleotide probes it will be necessary to clone and sequence the β globin genes. Since it has been found that particular mutations are linked to specific patterns of restriction fragment length polymorphisms of the β globin gene (haplotypes) it is helpful to carry out a preliminary haplotype analysis to determine which gene to study in detail (Orkin et al, 1982). It should be possible to facilitate the identification of point mutations by some of the recently developed methods for their rapid detection in genomic DNA (Myers et al, 1985). Once the common mutations in a given population have been identified an effort should be made to see if any of them could be identified directly using restriction enzymes or whether they are in linkage disequilibrium with particular restriction fragment length polymorphisms. But the most direct approach is to construct a

small number of oligonucleotide probes that are capable of diagnosing the bulk of cases in a particular population.

Currently, the use of oligonucleotide gene probes for diagnosis diseases from DNA obtained by chorion villus sampling is technically quite difficult and requires radioactive labeling of the probes. Recently however techniques have been developed for the rapid *in vitro* enzymatic amplification of genomic sequences (Saiki *et al*, 1985). By using this approach it may be possible to use a simple DNA dot blot analysis with diagnostic oligonucleotides rather than formal Southern blotting. If these new methods can be combined with non-radioactive labeling of probes the procedure will be even simpler and could be adapted for small rural hospital laboratories.

In summary, the global application of prenatal diagnosis to the major hemoglobin disorders will require considerable international cooperation and organization. A few central laboratories will have to be either set up *de novo*, or if already working in this field asked to provide a central diagnostic service. Their role will be to define the particular hemoglobin disorders and their molecular pathology for individual populations. At the same time local arrangements for screening and education should be developed. In addition it will be necessary to train a small number of obstetricians in each country in the techniques of chorion villus sampling. Once this has been done, and once the specific mutations that occur commonly in populations have been defined it should be possible to introduce simplified dot blot type analysis for prenatal diagnosis. Finally, it may be necessary for each country to develop perhaps one central diagnostic laboratory, rather like the blood group reference laboratories in the developed countries, that can deal with difficult cases in which oligonucleotide probe technology is not able to arrive at a diagnosis

and in which restriction fragment length polymorphism linkage analysis is required.

There has already been sufficient experience in countries such as Cyprus, Sardinia, Greece, Israel, and Italy, to suggest that prenatal diagnosis programs of this type are acceptable to the community and can be incorporated into current medical services. The fall in the birthrate of β thalassemia homozygotes in each of these populations over the last few years provides excellent evidence of what can be achieved (WHO, 1982; 1983).

The major question that remains is whether a simplified DNA technology such as that described above can be used in a large rural population, Thailand with its half million seriously affected children for example. Currently, a pilot program of this type is being set up in Thailand and if it is successful it will provide a very useful model system for many countries with large rural populations. The author's laboratory has analyzed the different mutations that occur in the Thai population and it has been estimated that two or three oligonucleotide probes would cover about 80% of 'at risk' pregnancies. A rural-hospital diagnostic service has been established and the ground seems to be set for such a pilot program to start within the next few years.

Strategies For The Control Of Other Monogenic Disorders

The approach to the population control of the hemoglobin disorders by carrier detection and fetal DNA analysis outlined in the previous section has general application to other common monogenic disorders. However, until more is known about their molecular heterogeneity it is difficult to provide an overall approach. However, several facts are now clear.

It seems unlikely that it will be possible to carry out population or maternal screening unless conditions are particularly common in a particular population. For example, it is clear that screening of 'at risk' population for the hemoglobin disorders is cost-effective. It will be similarly feasible to screen individual populations with an unusually high frequency for a particular genetic defect, Tay Sachs disease in Ashkenazi Jews for example. If it turns out that cystic fibrosis, with a carrier rate of 1 in 20 in populations of north European origin, is homogeneous at the molecular level it should be possible to develop oligonucleotide probe screening programs for all pregnant women. And it might be possible to include a number of other probes in a program of this type. As mentioned in the previous section it should soon be feasible to incorporate the screening programs into a simple dot blot approach to both screening and fetal DNA analysis.

It should be remembered however that many X-linked disorders such as Duchenne muscular dystrophy and hemophilia are the result of new mutations. There is increasing evidence that these disorders are highly heterogeneous at the molecular level. Thus screening for these conditions will have to be carried out only in families where there has been a previously affected child, and it seems unlikely, at least for the forseeable future, that it will be possible to carry out prenatal diagnosis by direct analysis; in all these conditions linkage analysis using restriction fragment length polymorphisms will have to be utilized. The only common X-linked disorder in the third world is glucose-6-phosphate dehydrogenase deficiency. Although this is not a disorder that would be considered for prenatal diagnosis, rapid carrier detection would be extremely helpful, particularly with the resurgence of malaria and the fact that individuals deficient in this enzyme are at high risk for hemolysis with antimalarial drugs.

Although there are many mutations that underlie this condition different populations seem to have a high frequency of the same mutation (see Luzzatto and Testa, 1978). Although there are relatively simple biochemical screening techniques available for this condition a rapid dot blot analysis would be extremely valuable. But the other X-linked disorders will have to be diagnosed in laboratories with the facilities for carrying out linkage analyses with restriction fragment length polymorphisms.

In developing these prenatal diagnosis programs it is clear that once particular mutations that underlie single gene disorders are known it should be possible to provide simple diagnostic kits that are feasible for relatively unsophisticated laboratories. But even when the majority of the genes for the important monogenic disorders have been isolated it will still be necessary to have some centralised laboratories that are capable of working out the individual mutations that are responsible for particular genetic disorders within different populations. Although it seems likely that simple kits for carrier detection or prenatal diagnosis will be produced commercially it is difficult to see at the current time where the funding will come from for the more sophisticated background studies required before it is possible to use diagnostic kits of this type in the field. Again, the thalassemias are a good example of some of the problems involved. The major research laboratories in the developed countries do not feel that it is appropriate for them to work out the mutations for these conditions in different populations in the third world and yet prevention programs cannot be started until this baseline information has been obtained. Who is going to do this work? As mentioned earlier, WHO designated and supported laboratories might be the obvious solution.

PRIORITIES FOR THE GLOBAL APPLICATION OF GENOME MAPPING

The frequency and prevalence of monogenic disorders in western societies was summarised earlier this report. Globally, the most important genetic disorders are the hemoglobinopathies and the common red cell enzyme defects. So far, very little is known about the world distribution of many of the other monogenic disorders with the exception of those shown in Table 2. Clearly for individual countries a high priority will have to be given to particular monogenic disorders that give rise to a major health problem, the porphyrias in South Africa and the adreno-genital syndrome in Eskimos for example. There are a number of serious genetic diseases which occur at a high frequency in other circumscribed regions, the dominantly inherited form of spino-cerebellar degeneration which is seen in certain parts of India for example.

From the point of view of world health the characterisation of susceptibility to common polygenic disorders is the other main priority. As mentioned earlier the whole question of genetic susceptibility to common infection has been largely ignored over the last few years. However, examples of striking variation in the distribution of the class I and class II alleles of the HLA genes in conditions such as malaria and the different subtypes of leprosy point to the extreme importance of attempting to define susceptibility loci of this kind. Similarly, the problem of genetic susceptibility to heart disease and hypertension is becoming an equally important problem for the third world; recent surveys indicate that these conditions are taking the place of malnutrition and infectious disease, with improved standards of living.

As is the case for prenatal diagnosis programs any practical fallout from the gene mapping program will have to be developed to a

stage where the technology is simple and applicable to large populations. In practice, this will require the development of techniques for defining individual mutations that are applicable in the field. Thus it will be essential to continue to simplify analytical techniques. Currently there are many promising leads in this direction. For example, the further development of the polymerase chain reaction and the use of specific oligonucleotide probes together with the development of more efficient methods for non-radioactive labeling of the probes should make it possible to develop simple kits for analysing single mutant genes. This type of technology has broad application for the detection of single gene disorders but also for defining particular susceptibility alleles for some of the common polygenic conditions.

At this early stage it is difficult to define precisely the organisational problems that may arise in attempting to move some of this high technology from western laboratories to where it is required in the third world. It seems likely that most of the developmental work will have to be done in the west and that a series of designated training laboratories will have to be developed, possibly by the WHO, for teaching the application of these techniques. As gene probes become more widely used it may be more practical for strong links between the academic departments and industry to develop for the dissemination of simplified kits for gene analysis. There is no doubt that some kind of centralised database will be required and again a central body like the WHO might be best equipped to disseminate this type of information.

As mentioned previously, the hemoglobin disorders provide a very good example of the position that might be reached with other single gene and polygenic conditions over the next few years, once the human

genome has been mapped. In this case we already have a clear and detailed map of the regions of the genome involved and know a great about the mutations that cause these conditions. But it is still not clear who is going to carry out the next step i.e. to analyse particular mutations within individual populations as a basis for setting up screening and prenatal diagnosis programs. Clearly a central laboratory, or several laboratories should do this but it is not clear where the funding will come from, who will organise the whole thing, or how this technology will actually be disseminated to the third world. Currently, this is being carried out on individual bases between third world workers and those in western laboratories; this is an ad hoc arrangement and large areas of the world in which this technology could be usefully supplied do not even know that it exists. Although the WHO have made a start with this type of organisational difficulty, even in the case of very well defined disorders like the thalassemias the position is far from satisfactory. And it is quite clear that as third world countries become more affluent, and genetic disease becomes more of a problem, the demand for this kind of information will grow.

CONCLUSIONS

Genetic diseases are an important cause of ill health in the developed world and will become an increasingly serious problem in the developing world as social conditions improve and the high mortality in early life due to infectious diseases and malnutrition is controlled. Although the hemoglobin disorders have been the prime model for studying the global problems of the control of genetic disease there is increasing evidence that other genetic disorders occur commonly in particular populations. The development of a linkage map, and ultimately the complete sequence of the human genome

will greatly facilitate the isolation of the genes involved in monogenic disorders and help to define the polygenic basis for many common disorders of the developed and developing world populations. There is already enough evidence from the globin disorders to suggest that once individual mutations are known a very simple technology can be used for carrier detection and prenatal diagnosis using DNA analysis. However, it is clear that if this technology is to be used widely, particularly in the developing world, a considerable amount of more sophisticated backup information will be required, particularly working out the particular mutations involved in genetic diseases in individual racial groups. Thus, as a backup to the gene sequencing program some kind of central sequencing facilities will be required to analyse mutant genes in different populations. Only if this is done will the data derived from the genome sequencing programs be made available for the control of genetic disease throughout the world. Unless some of these problems are considered at an early stage in the development of the sequencing program there will be an enormous delay before the benefits that will undoubtedly follow are translated into day to day clinical application.

REFERENCES

Anderton, B.H. (1987) Progress in molecular pathology. Nature, 325, 658-659.

Bishop, M.J. (1987) The molecular genetics of cancer. Science, 235, 305-311.

Bodmer, W.F. (1986) Human gentics: the molecular challenge. In: Cold Spring Harbor Symposia on Quantitative Biology, 51, 1-13.

Bowden, D.K., Hill, A.V.S., Higgs, D.R. and Weatherall, D.J. (1985) Relative roles of genetic factors, dietary deficiency, and infection in anaemia in Vanuatu, South-West Pacific. Lancet, ii, 1025-1028.

Brown, M.S. and Goldstein, J.L. (1986) A receptor-mediated pathway for cholesterol homeostasis. Science, 232, 34-47.

Bunn, H.F. and Forget, B.G. (1986) Hemoglobin: Molecular, Genetic and Clinical Aspects W.B. Saunders, Philadelphia.

Deeb, S., Failor, A., Brown, B.G., Brunzell, J.D., Albers, J.J. and Motulsky A.G. (1986) Molecular genetics of apolipoproteins and coronary heart disease. In: Cold Spring Harbor Symposia on Quantitative Biology, 51, 403-409.

Flint, J., Hill, A.V.S., Bowden, D.K., Oppenheimer, S.J., Sill, P.R., Serjeantson, S.W., Bana-Koiri, J., Bhatia, K., Alpers, M.P., Boyce, A.J., Weatherall, D.J. and Clegg, J.B. (1986) High frequencies of α thalassaemia are the result of natural selection by malaria. Nature, 321, 744-749.

Jeffreys, A.J., Wilson, V. and Thein, S.L. (1985) Hypervariable "minisatellite" regions in human DNA. Nature, 314, 67-73.

Kang, J., Lemaire, H-G., Unterbeck, A., Salbaum, J.M., Masters, C.L., Grzeschik, K-H., Multhaup, G., Beyreuther, K. and Muller-Hill, B. (1987) The precursor of Alzheimer's disease amyloid A4 protein resembles a cell-surface receptor. Nature, 325, 733-736.

REFERENCES

Kar, B.C., Satapathy, R.K., Kulozik, A.E., Kulozik, M., Sirr, S., Serjeant, B.E. and Serjeant, G.R. (1986) Sickle cell disease in Orissa State, India. Lancet, ii, 1198-1201.

Kidson, C., Lamont, G., Saul, A. and Nurse, G.T. (1981) Ovalocytic erythrocytes from Melanesians are resistant to invasion by malaria parasites in culture. Proc. Natl. Acad. Sci. U.S.A. 78, 5829-5832.

King, M.H. (1986) Medicine in an unjust world. The Oxford Textbook of Medicine 2nd Edn (D. J. Weatherall, J. G. G. Ledingham and D. Warrell), pp. 3.1-3.10. Oxford University Press, Oxford.

Livingstone, F.A. (1967) Abnormal hemoglobins in human populations. Aldine, Chicago.

Livingstone, F.B. (1985) Frequencies of Hemoglobin Variants Oxford University Press, New York.

Luzzatto, L. and Testa, U. (1978) Human erythrocyte glucose- 6-phosphate dehydrogenase: structure and function in normal and mutant subjects. Current Topics in Hematology , 1, 1-70.

Moore, G.E., Ivens, A., Chambers, J., Farrall, M., Williamson, R., Page, D.C., Bjornsson, A., Arnason, A. and Jensson, O. (1987) Linkage of an X-chromsome cleft palate gene. Nature, 326, 91-92.

Myers, R.M., Lumelsky, N., Lerman, L.S. and Maniatis, T. (1985) Detection of single base substitutions in total genomic DNA. Nature, 313, 495-498.

Orkin, S.H. and Kazazian, H.H. (1984) The mutation and polymorphism of the human β-globin gene and its surrounding DNA. Annu. Rev. Genet, 18, 131-171.

Orkin, S., Kazazian, H.H., Antonarakis, S.E., Goff, S.C., Boehm, C.D., Sexton, J.P., Waber, P.G. and Giardina, P.J.V. (1982) Linkage of β-thalassaemia mutations and β-globin gene polymorphisms with DNA polymorphisms in human β-globin gene cluster. Nature, 296, 627-

631.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N. (1985) Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science, 230, 1350-1354.

Scriver, C.R., Neal, J.L., Saginur, R. and Clow, A. (1973) The frequency of genetic disease and congenital malformation among patients in a pediatric hospital. Canad. Med. Assoc. J, 108, 1111-1115.

Stamatoyannopoulos, G. (1973) Problems of screening and counseling in the hemoglobinopathies. Proc. IV Internat. Conf. Birth Defects, Vienna. p.268.

United Nations Secretariat. (1982) Long range global population projections, as assessed in 1980. Population Bulletin of the United Nations, 14, 17-30.

Vogel, F. and Motulsky, A.G. (1982) Human Gentics. Problems and Approaches Springer-Verlag, Berlin Heidelberg.

Weatherall, D.J. (1985a) The New Genetics and Clinical Practice 2nd Edn Oxford University Press, Oxford.

Weatherall, D.J. (1985b) Prenatal diagnosis of inherited blood diseases. Clinics in Haematology, 14, 747-774.

Weatherall, D.J. (1988) The new genetics. In: New Prospects for Medicine. Ed. Austyn, J.M. Oxford University Press. In press.

Weatherall, D.J., Clegg, J.B., Higgs, D.R. and Wood, W.G. (1988) The hemoglobinopathies. The Metabolic Basis of Inherited Disease 6th Edn (C. R. Scriver, A. L. Beaudet, W. S. Sly and D. Valle), McGraw-Hill Book Co, New York. In press.

Weatherall, D.J. and Clegg, J.B. (1981) The Thalassaemia Syndromes 3rd Edn Blackwell Scientific Publications, Oxford.

REFERENCES

Weatherall, D.J., Higgs, D.R., Bunch, C., Old, J.M., Hunt, D.M., Pressley, L., Clegg, J.B., Bethlenfalvay, N., Sjolin, S., Koler, R.C., Francis, J.L. and Bebbington, D. (1981) Hemoglobin H disease and mental retardation: a new syndrome or a remarkable coincidence? New Eng. J. Med, 305, 607-612.

Weatherall, D.J., Old, J.M., Thein, S.L. and Wainscoat, J.S. (1987) The role of cloned genes in the prevention of genetic disease. Proc. Roy. Soc. Med. In press.

WHO WORKING GROUP. (1982) Hereditary anaemias: genetic basis, clinical features, diagnosis and treatment. Bull WHO, 60, 643-660.

WHO Working Group. (1983) Community control of hereditary anaemias. Bull WHO, 60, 63-80.

IN SEARCH OF THE "ULTIMATE MAP" OF THE HUMAN GENOME

THE JAPANESE EFFORTS[1]

Akihiro Yoshikawa, Ph.D.
Berkeley Roundtable on the International Economy (BRIE)
215 Moses Hall
University of California, Berkeley
Berkeley, CA 94720
(415) 642-3067

I. INTRODUCTION

The goal of this study is to describe and analyze the Japanese efforts to sequence DNA, and to assess their impacts upon the US.[2] Section 2 provides a short history of the Japanese sequencing efforts. Also discussed are the roles of government ministries, business and academic advisory committees in the formation and implementation of policy. In Section 3, the emphasis on developing automated technologies in sequencing will be analyzed in detail. Plans to establish other related infrastructures, i.e. data banks and repositories, are also discussed here. In both Section 2 and 3, possible future impacts of government involvement in sequencing efforts on the commercialization of technologies are explored.

---

[1] This was prepared for the Office of Technology Assessment (OTA), U.S. Congress. The main part of this study is based upon interviews I conducted in Japan between April 27 and May 1, and also between May 27 and June 4, 1987. I am greatly indebted to reviewers of my first draft.

[2] For general review of Japanese policies to promote biotechnology, see Gary Saxonhouse, "Industrial Policy and Factor Market," in Hugh Patrick ed., Japan's High Technology Industries (Seattle: University of Washington Press, 1986), and Akihiro Yoshikawa, The Japanese Challenge in Biotechnology: Industrial Policy (Berkeley, BRIE Working Paper, University of California, 1987).

Some technologies have been already developed through the Japanese DNA sequencing project and have already been commercialized. For example, Seiko has developed a microchemical robot for automation of the chemical and the enzymatic sequencing reactions, and Toyo Soda has developed DNA extractors with a subsidy from the government.

Prime Minister Nakasone introduced the Human Frontier Science Program (HFSP) at the 1987 Summit in Italy. Though the program is not yet well defined, it is likely to affect the implementation of the sequencing effort since DNA sequencing is regarded as one of the most important technologies related to the Human Frontier Science Program. Therefore, Section 4 will discuss the Human Frontier Science Program--its goals, impacts on the sequencing efforts, and also its shortcomings. Section 5 will discuss whether any international cooperation is possible. If so, in what respect? What are the difficulties? What are possible forms of international cooperations, i.e. work sharing or establishment of research funds, for the sequencing efforts? What is the future of the Human Frontier Science Program?

Additionally, one must remember that both the sequencing project and the Human Frontier Science Program are currently still being studied by the government.

II. JAPANESE EFFORTS IN DNA SEQUENCING

A. Short History of the Japanese Sequencing Efforts

An initial effort to promote the mass-reading of DNA-base sequence was started in 1981. The program was entitled, "Extraction, Analysis, and Synthesis of DNA," and was lead by Professor Akiyoshi Wada of the University of Tokyo. The aim of the program was "to encourage DNA study in Japan by organizing the resources available in universities, private companies, and government research institutes."[3] Since the program tried to coordinate universities and government with private firms, the research "ought to have applications broader than of a purely academic nature."[4] Professor Wada described the emphasis of the program.

> ...the analysis section of the program has tried to reduce the burden of time demanded of researchers working on the analysis of DNA base-sequence by developing automatic machinery. At the same time, the aim was to maximally utilize the potential technological resources of non-biological companies, because the future development of pure and applied biology is believed to depend much on physical techniques, including those of electronics, computers, robotics and materials sciences. We started by substituting robotic techniques and mass production machines for established processes, rather than developing new processes. It does not seem wise to simultaneously change two factors which are interdependent, and we felt that highest priority should be given to automation.

---

[3] Wada, 1986, pp. 16.

[4] Ibid.

The program was financed by the Special Coordination Fund for the Promotion of Science and Technology (Kagakugijutsu Shinkochosei-hi). The project was one of the nine projects designated by the Science and Technology Agency (STA) as Special Priority Fields of Research. (See List 1 below.) The sequencing effort was annexed to the Research on Development of Common Basic Technologies for Cancer Research in 1984. Currently more detailed discussions concerning goals and methods are conducted under the auspices of the STA, in the Council for Aeronautics, Electronics and other Advanced Technologies.

List 1. STA's Special Priority Fields of Research

1	Research project for development of DNA extraction, analysis and synthesis technologies
2	General research on safety of rDNA technologies
3	General research relating to utilizing rDNA technologies
4	Research for development of technologies for the analysis and utilization of biomembrance functions
5	Research relating to development and cultivation of experimental animals
6	Research relating to large-size super high-pressure generation systems
7	Research for development of basic technologies relating to superconductivity and cryogenic temperatures
8	Research relating to surface and boundary control technologies for developing high-performance materials
9	Research for developing new materials by means of gravity-free environment

Science and Technology in Japan, pp.51.

B. Policy Formation and Implementation: Special Coordination Fund and Research Development Corporation

The Special Coordination Fund for the Promotion of Science and Technology was established in 1981 in order to promote science and technology in Japan. The general use of the fund is set by the Policy Committee (Seisaku Iinkai) of the Council for Science and Technology (Kagaku Gijutsu Kaigi). It defines Japan's long term goals for science and technology development. In 1986, the Policy Subcommittee consisted of eleven members. It included top executives from large Japanese corporations, such as NEC and Mitsubishi Chemical.

Whereas Special Coordination Funds provided an incentive for basic research for new technologies, the Research Development Corporation (Shingijutsu Kaihatsu Jigyo-dan) played an active role in the commercial development of technologies. It is a government-funded Public Corporation (jigyodan), and was established in 1961 in order to promote commercial use of government developed technologies.[5] It was designed to encourage uses of technologies which were not likely to be utilized commercially in the near future, without the help of the government. It collects information

[5] Public Corporations (Tokushu Hojin) are corporations established by the Japanese government under a special law. Public corporations are formed when a specific activity can be more efficiently operated by a separate entity, rather than directly by the government. As of January 1987, there are 87 public corporations, created for various purposes. Research Development Corporation of Japan and Japan Information Center of Science and Technology (JICST), both under the auspices of STA, are examples of such public corporations. Each public corporation has a board, i.e. a president and vice presidents, who are appointed by the relevant ministers.

on the research findings of universities and research institutes, and organizes cooperative development projects with private companies. Research Development Corporation selects private companies to conduct specific projects in cooperation with their original inventors. In the Research Development Corporation, there is a Development Council which assists and submits recommendations to the board. On April 1, 1986, the Development Council consists of 15 members from academic and business circles.[6]

If a development is judged to be successful, the contractor is authorized to commercialize the new technology under a renewed contract, and expenses will be financed by Research Development Corporation. If such a development fails, the private firm does not have any obligation to pay back the funds received.

In the actual commercialization of the technology developed during such a project, the priority for commercialization is given to the company that had been contracted by the Research Development Corp. Depending on who was the original inventor, and whether or not the original inventor's research was funded by the government, its patent rights may be shared by the government, Research Development Corp. and a private contracting firm. If a firm not contracted by Research Development Corp. wishes to commercialize the technology, Research Development Corp. "may license this company to do so after considering the market requirements and licensing

---

[6] Members of the Development Council include Nobuto Ohta (President, Oji-Yuka Synthetic Paper Co.), Kin'ichi Nohga (Statutory Auditor, Toshiba Corporation), Sadataka Mukoyama (Mitsui Interbusiness Research Institute), and Shigeya Yoshise (Governor, The Japan Development Bank).

conditions."[7] One such consideration is whether the contractor could obtain favorable market shares for the product.[8] Research Development Corporation has consigned projects concerning development of interferon at Green Cross and Toray Industries in the past.

After a prototype was developed with funds made available from the Special Coordination Fund for the Promotion of Science and Technology, Research Development Corporation assisted Seiko Instruments and Electronics Ltd. to commercialize its microchemical manipulator to automate the Maxam-Gilbert sequencing reactions. (More details on the Japanese efforts in developing hardware will be discussed below.)

C. Science Policy Formation: Public Disucssion

One notable difference between the U.S. and Japanese process of science policy formation is that whereas in the U.S. there are extensive public discussions among various experts, even in initial stages of policy discussions, there are no such public debates in Japan. Often policies are formulated among the "inner cabinet" of the Council for Science and Technology.[9] The Council for Science and Technology, which establishes the general framework of science

---

[7] Research Development Corp., pp. 5.

[8] The government will allow the contracting firm a sufficient amount of time so that it can establish its dominance in the market place.

[9] The policies emerging from the Council are rather general, and specific details are determined during the subsequent policy implementation process.

policy in Japan, is chaired by the Prime Minister. Prime Minister Nakasone has been one of the most active advocates of biotechnology in Japan, and it is reasonable to assume that the publicity and momentum that both the sequencing efforts and the HFSP have enjoyed are at least partially due to the Prime Minister's personal involvement in the life sciences.

However, as is typical of policy debates in Japan, there was no serious public discussion concerning whether or not to support the idea to develop automated technologies for mass sequencing of DNA. A quiet consensus was achieved at least upon developing automated technologies necessary for mass sequencing of DNA. It was assumed that development of automated technologies is useful regardless of whether the entire human genome is being sequenced. More careful discussion concerning goals and methods involved in the Japanese efforts to sequence DNA is currently being discussed in the Human Genome Sequencing Subcommittee at the Council for Aeronautics, Electronics and Other Advanced Technologies (CAEOAT).

D. Biotechnology Committee of CAEOAT

25 members of the Council for Aeronautics, Electronics and Other Advanced Technologies, as well as members of its Biotechnology Committee and 15 members of the Human Genome Sequencing Subcommittee are all selected by the Director-General of the Science and Technology Agency. See Figure A for the structure of the Council.

Some members in the Council and its Biotechnology Committee are top executives from Japanese private companies, though all

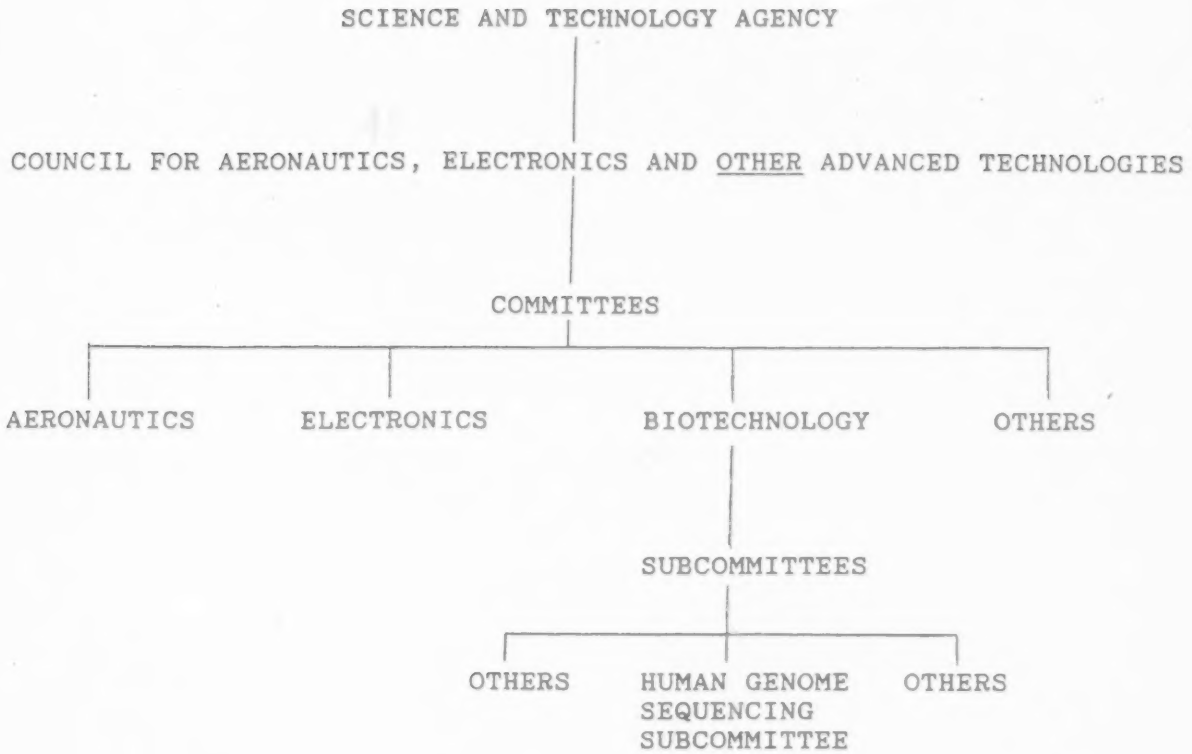
members of the subcommittee are scientists from universities and public research institutes.[10] However a leading member of the subcommittee insisted that the actual decision making concerning Japan's sequencing effort will be made by the 15 scientists on the subcommittee, but not members in the Biotechnology Committee or the Council itself. The Council plans to make its final decision by March 31, 1988. It is expected that the basic agreement concerning the final proposal will be achieved within the 1987 calendar year.

The Science and Technology Agency functions as the coordinator of the Fund. The Special Coordination Fund provides funds for the sequencing efforts. As noted below, Hitachi developed an advanced DNA sequencer utilizing the Special Coordination Fund. It is important to note that so far STA, not MITI, is the prime governmental body in the Japanese sequencing efforts.

---

[10] Members of CAEOAT (total of 25) includes Y. Iida (President, Mitsubishi Heavy), S. Iwasaki (Director General, NTT), M. Goto (Vice President, Research Development Corporation), K. Morita (Managing Director, Takeda Chemical), and H. Watanabe (Vice President, Hitachi).

FIGURE A.


SCIENCE AND TECHNOLOGY AGENCY
COUNCIL FOR AERONAUTICS, ELECTRONICS AND OTHER ADVANCED TECHNOLOGIES
COMMITTEES
AERONAUTICS
ELECTRONICS
BIOTECHNOLOGY
OTHERS
SUBCOMMITTEES
OTHERS
HUMAN GENOME SEQUENCING SUBCOMMITTEE
OTHERS

E. Okayama Conference

A workshop entitled International Workshop on Automatic and High-Speed DNA-Based Sequencing was held in Okayama (Okayama Conference), Japan, between July 7th and 9th. More than fifty persons, including twenty from abroad (West Germany, U.K., and U.S.A) participated in the meeting. The general purpose of the meeting was to discuss the current status of the automation and acceleration of DNA-base sequencing processes in the world and the strategy for establishing the total analysis of the human genome.

There were several presentations by Japanese companies (Seiko Instruments, Hitachi, and Fuji) who participated in the STA's sequencing program. Companies' presentations primarily reflected their development of automated sequencing related machines utilizing Special Coordination Funds.[11] The conference was at least partially motivated by the STA's desire to present the achievements of its Special Coordination Funds. It should be noted that such a demonstration was important in order to justify financial support to corporations and to secure its budget for the future.

Some Japanese participants, however, expressed some doubt toward the Japanese sequencing project led by STA.[12] They pointed to the lack of open debates concerning the DNA sequencing in Japan

---

[11] Seiko Instruments presented a paper entitled "Automation of the M13/dideoxy Method and the Gel-Pattern Reading," and Hitachi presented "Real Time Detection of Fluorescence Labeled DNA Fragments During Electrophoresis."

[12] It should also be mentioned that there was a certain bias concerning the Japanese participants, since many were either from companies involved in the STA project, or scientists with STA connections.

and also expressed their reservations toward the STA-led project. One Japanese scientist maintained that the STA project didn't generate active support from other Japanese scientists. One reason is that many Japanese scientists have closer ties with the Ministry of Education (MOE) rather than with the STA. Those scientists were not well informed about the STA project and were not especially supportive of such an attempt.[13]

F. Re-Assessment of Japanese Sequencing Project

Lastly, it should be mentioned that there are some overly optimistic expectations toward the Japanese ability to sequence DNA, particularly as reported in the US media. For example, Science (2/3/1987, pp. 31) reported that the Japanese super-sequencer would be expected to be up and running in one year. However, Professor Wada insists that these reports do not reflect the present level of automated high-speed DNA sequencing in Japan, and that much more work will be needed to make the system operable.

---

[13] Though I suspected that the conference would probably set a pace and direction concerning the Japanese sequencing effort, there was no clear indication of either at the conference. Apparently the 15 member subcommittee had not yet developed a concrete recommendation. Also missing from the meeting was any indication of the future relationship between the sequencing effort and the Human Frontier Science Program (HFSP).

III. AUTOMATED TECHNOLOGIES

A. Japanese Emphasis on the Development of Hardware

The Japanese strategy to create an efficient method for high speed DNA sequencing is to develop automated tools for various steps involved in sequencing, and to link these steps effectively so that the process of sequencing can be smoothly linked. Thus development of tools used in various steps and their compatibility are the essential factors in the Japanese effort to realize efficient mass sequencing flow.

Prof. Wada, who has been the leading (and most influential) scientist in the Japanese sequencing project, stated:

> There are two aspects to improving the sequencing rate: speeding-up each machine used in processing (chemical processor, electrophoresis, autoradiograph, etc.), and careful organizing of not only the processing but also each unit involved so that they operate continuously.[14]

It should be noted, however, that Prof. Wada's strategy will not eliminate the need for skilled labor inputs to DNA sequencing; rather the involvement of skilled workers will be crucial.

> We must emphasize that our goal is not that of a fully robotic system, but of a DNA sequencing line capable of yielding the sequence of 1Mb per day in which skilled human operators would play a crucial part in monitoring the accuracy of the results.[15]

---

[14] A. Wada & E. Soeda, "Strategy for Building an Automatic and High Speed DNA Sequencing System," in the Proceedings of the 4th Congress of the Federation of Asian and Oceanian Biochemists, 1986.

[15] A. Wada, "Automated high-speed DNA sequencing," in Nature, 26, February, 1987.

Prof. Wada has stressed the parallel between the mass reading of DNA sequencing and other areas where mass production systems have been applied, i.e. automobile and consumer electronics industries.[16] However, it should be noted that this does not mean that Japan is not trying to develop "new" methods in sequencing.

> Major individual techniques, for instance a high resolution electrophoresis device to separate and analyze simultaneously extremely small DNA fragments and an analytical method using fluorescent dye instead of radioisotopes, should also be improved on the basis of new principles.[17]

On June 6, 1987, Hitachi announced that the firm developed a DNA sequenator based on a "new" method, which utilizes techniques using fluorescent dyes. Its development was supported by the government. (More details are discussed below.)

B. Overview of Company Projects

Various Japanese companies such as Seiko Electronics and Instruments, Fuji Photo Film, Toyo Soda, Hitachi, and Mitsui Knowledge Industry, have received governmental support. Seiko developed microchemical robots to carry out a series of chemical manipulations according to the Maxam-Gilbert protocol. Its development was supported by the Research Development Corporation.

---

[16] A. Wada, June 1987.

[17] Current Developments in Research in Relation to the Principal Biological Functions and Future Focal Points, pp. 57-58.

(The prototype was developed with funds available from the Special Coordination Fund.) The firm also developed, in collaboration with scientists at the Institute of Physical and Chemical Research, a prototype microchemical robot based on Sanger's M13 dideoxy. Similarly with support from the government, Fuji Photo Film developed a machine for the manufacture of prepackaged acrylamide gel film. Both Hitachi Software and Seiko independently developed automatic scanners of autoradiograph. Mitsui Knowledge Industry developed computer software for analyzing DNA base sequences.

Hitachi recently developed DNA sequencer based on Sanger's method.[18] The machine is very similar (or essentially the same) as the one introduced by Applied Biosystems. Its development was supported by STA's Research on Development of Common Basic Technologies for Cancer Research, and the fund has been provided by the Special Coordination Fund for the Promotion of Science and Technology. It is reported that the machine can operate at least twice as fast as the one developed by applied Biosystems based on the same protocol.[19]

It is apparent that the Japanese government has been successful in encouraging the development of an efficient mass DNA sequencing flow by utilizing governmental support or by providing an incentive (targeting).[20]

---

[18] Nihon Keizai, 6/6/1987.

[19] Nihon Keizai, 6/6/1987.

[20] Whether such governmental support can insure commercial success of Japanese firms is not clear. It is important to realize that a development of a prototype itself doesn't guarantee success

C. Caution in Interpretation of Japanese Efforts

Though it seems that the Japanese government is supporting the development of hardware (such as DNA sequencing machine), one should also realize that its actual financial support to companies is rather small. Additionally it should be noted that the U.S. Applied Biosystems' equipment was developed in collaboration with Cal Tech scientists who were supported by a federal grant.[21]

Japanese executives from companies I interviewed also complained that the financial support from the STA was terribly inadequate, and that the government was taking too much credit for the development. One observer maintained that the government had been getting too much credit than they actually deserved, and MITI and STA tended to support the projects which were already going well at academic and national laboratories.[22] It is argued that many of these commercializations could be realized without government help. As mentioned above, the Okayama Conference was at least partially motivated by the STA's desire to publicize the achievements of its Special Coordination Funds.

Both STA and MITI publicize their "accomplishments" widely in Japan mainly in order to gain greater influence in domestic politics and in the business community. Recently such public relation

---

in the market place. American executives may insist that competitive pressure in the market, not coordinated efforts led by the government, is the true driving force of innovation.

[21] This point was raised by one of the reviewers of the first draft.

[22] This point was raised by a reviewer.

attempts have been closely watched by foreign press and scholars. Many foreign observers were convinced by these public relation attempts that the Japanese government played a very large role in the development of commercialization. It may be rather unfortunate for Japanese companies and also for Japan in general in terms of international economic relations, if Japanese firms could have successfully commercialized these technologies by themselves without government support.

D. Possible Conflict of Interests

Interviews in Japan suggest that Japanese bureaucrats and some scientists intend to use Japanese hardware for Japanese sequencing efforts, even if US machines are currently available. As we said before, with funds from the STA's Special Coordination Fund, Hitachi recently developed a DNA sequenator which is similar to the one marketed by an American firm, Applied Biosystems. Applied Biosystem representatives and American policy makers, of course, feel uneasy toward such Japanese policy intended to promote its own machine.

The first explanation for Japanese government support of hardware development comes from the Japanese government bureaucrats, particularly those at STA. They have stated that Japan is developing the machine because the U.S. firm's equipment is not satisfactory in terms of cost, performance or quality.[23] Similar

---

[23] The statement was directed not only to laboratory equipment in general but also specifically to the U.S. made DNA sequencing machine, even though Applied Biosystems equipment are very popular today in Japan.

views have been expressed in the Japanese literature, including some publications by the Japanese Bioindustry Development Center (BIDEC), which is an industrial association of Japanese biotechnology related companies.

The second explanation for government support of hardware development was expressed by Professor Wada, a university professor who is the leading researcher in this field. He maintains that the coordination of various machines utilizing "conventional" technologies is more suitable to mass production than the development of a "new fully robotic" method, such as that being developed by Applied Biosystems. Professor Wada has been the key scientist in the development of the Seiko's chemical robot, Fuji's gel machine, and other equipments utilizing "conventional technologies." As a scientist, he is convinced that well coordinated use of conventional machines is, at least for time being, better than developing new machines.

In conclusion, the Japanese desire to develop Japanese made machines in sequencing and other related projects (automation in the construction of ordered DNA clones is now under study) may develop into a crucial issue between the U.S. and Japan. Regardless of whether Hitachi's development of a prototype realizes commercial success (over American firms), it may raise a serious question of what the Japanese government should do when the U.S. in fact tries to cooperate with Japan in the sequencing effort.

E. Repository/Data Bank

Of equal importance as machines in the sequencing project, is the establishment of a well organized repository and data bank. In the US, the American Type Culture Collection (ATCC), a non-profit organization, maintains various collections of strains of algae, bacteria, fungi, plant and animal viruses, protozoa and cell lines, as well as a collection of plasmids, rDNA vectors and hybridomas.[24] In DNA sequencing, GenBank has been playing a very important role though its capacity could not satisfy the current flow of DNA sequencing data. This section will explore the Japanese discussions concerning the development of DNA repositories, data banks, and other plans for storing DNA clones and other materials involved.

Current conditions concerning repositories and data banks in Japan are far from satisfactory to support its R&D. Many firms and government ministries have maintained their own data banks and repositories, however, none of them could be compared to the ATCC or GenBank. The government is now trying to establish its own repositories/data bank to provide the necessary infrastructure for its sequencing effort.[25] On March 28, 1986, the Japanese Cabinet declared that the development of a scientific information network as well as repositories/data banks for biological materials were essential for scientific development.[26] MITI's budget for 1986 for

---

[24] Elkingon, pp. 189.

[25] STA, 1986, pp. 121-130.

[26] Ibid., pp. 296.

repository/data bank related areas was 723 million yen.[27] The total expense for the repository/data bank would be greater since funds to finance the Protein Engineering Research Institute (PERI) and its protein data bank are not included in the above figure. STA's budget for cell-gene repository was 226 million yen.

Several ministries are now trying to establish their own repository/data bank. The Ministry of Health and Welfare established the Research Resource Bank, which collects DNA data mainly for cancer research. STA is now trying to develop its own gene bank at the Institute of Physical and Chemical Research. The Ministry of Agriculture, Forestry and Fisheries established its own gene bank and data bank for plant DNA sequencing. MITI is also trying to establish its protein data bank in Japan at the Protein Engineering Research Institute (PERI).[28]

Protein Engineering Research Institute is now under construction by MITI and MPT, and its membership is shared by five companies: Toray, Mitsubishi Chemical, Kyowa Hakko, Takeda Chemical, and Toa Nenryo Kagaku. MITI and the Ministry of Post and Telecommunication (MPT) financed 70% of the capital, and the remaining was shared by the five companies.[29] Protein Engineering Research Institute is planning to develop a data base for DNA,

---

[27] BIDEC News, no. 43.

[28] Nikkei Sangyo Shinbun, pp. 187.

[29] These funds were made available through the sale of Nippon Telephone and Telegraph (NTT) stock. Nine additional companies (Kirin Beer, Suntory, Nippon DEC, Nippon Roche, etc.) joined the PERI later.

amino acid, peptide, and protein, and necessary software for effective utilization and analysis of this data.

DNA DATA Bank of Japan (DDBJ) was established at the National Institute of Genetics in 1984. However due to the lack of both manpower and a good computer system, its operation was minimal during the first two years. Its activities were limited to relaying data obtained from Gen Bank and EMBL to researchers. Recently, it began to input its own collection of data (mainly from publications in Japan's academic journals). The National Institute of Genetics is under the authority of the Ministry of Education (MOE).

F. International Cooperation in Repository/Data Base

Some scientists have pointed out that international cooperation is both desireable and feasible in the establishment of a repository and data base. However there may be a problem concerning the issue of cooperation in connection with repositories of DNA clones, DNA probes, DNA cloning vectors, cells, etc., since these may involve information closely related to commercial development. Several important questions arise in considering this problem.

First of all, what is the Japanese policy in depositing clones, probes, or cell lines to be patented? Currently the Fermentation Research Institute functions as the authorized depository for patented microorganisms in Japan. However, areas which are covered by patent are not yet clear, especially with regard to sequencing data.[30] At this moment, there is not sufficient discussion of this

---

[30] This is also true for the U.S.

legal issue in Japan.

The last question concerning international cooperation in the development of repositories is whether Japan has any specific plans for financial and logistical support of the existing international data bases for proteins and DNA, i.e. GenBank and EMBL. This would be the simplest and perhaps the most efficient way to promote international cooperation. Though one STA official maintained that Japan was interested in establishing its own data base and repositories and not merely supporting the already existing EMBL or GenBank, Japanese scientists in general expressed a more possitive view toward such Japanese supports for existing data bases abroad. More active open debate with Japanese scientists concerning this issue is recommended.

IV. HUMAN FRONTIER SCIENCE PROGRAM

A. Introduction

Faced with criticism from abroad that Japan has been a "free rider" in technology development, as well as the problem of increasing trade friction, the government of Japan is preparing a seemingly unique scientific plan called the Human Frontier Science Program (HFSP). The original ideas concerning the HTSP were developed by the Agency of Industrial Science and Technology at the Ministry of International Trade and Industry.

Unlike policies intended to achieve successful commercialization in Japanese industries, the HFSP was designed to

solve "global problems facing the human race." A memo prepared by the Agency of Industrial Science and Technology states:

> We propose to promote basic research on a world scale-- the "Human Frontier Science Program." This program aims to elucidate various biological functions and to apply these findings in the future. Biological functions can provide a vital key to developing new modes of science and technology...The Human Frontier Science Program is envisaged as an international basic research program, aimed at elucidating various biological functions and applying these findings in the future." (MITI, January 1987)

The HFSP, according to the government, is Japan's effort to contribute to basic scientific knowledge in the field of biological functions. For example,

A)the changing functions of biomaterials and energy: the functions of enzymes in digestion and absorption, and photosynthesis.

B)the functions of the conversion of information by organisms: mechanisms by which the brain can think and memorize, and sensory mechanisms.

Currently, the government is conducting a feasibility study regarding desireable topics of study and possible organizational frameworks. Figure B shows a proposed framework of the Program, and Figure C shows its proposed research subjects. The 1987 budget for the HFSP feasibility study is shown in Table A, and members of the Feasibility Committees are listed in Appendix A.

Though at present there is no concrete plan concerning its actual implementation, one of the Japanese proposals calls for the

establishment of a "Center for the Promotion of the Human Frontier Science Program." This center would, among other tasks, distribute research grants to research institutes throughout the world.

B. Chronology

According to a MITI source an initial study concerning HFSP was begun in September 1985 by the Working Group for Technology and International Exchange, an advisory committee responsible to the Director-General of the Agency of Industrial Science and Technology (AIST). In February 1986 a report (Proposal for Technological Development and International Exchange in the Years Approaching the 21st Century) was released. The report proposed to conduct a large scale basic research program focused on the study of biological functions. The proposal stressed the importance of international cooperation and was named the Human Frontier Program. Faced with trade frictions abroad and problems concerning structural readjustment of Japanese industries, this approach to international cooperation was supported in subsequent government advisory groups.

Prime Minister Nakasone has also supported the development of the HFSP. With his support, the Committee on Policy Matters of the Council for Science and Technology started to study means of implementation on May 27, 1986. In August of 1986, the first interministerial liaison conference concerning the HFSP was held. MITI, STA, Ministry of Foreign Affairs, Ministry of Education, Ministry of Health and Welfare, Ministry of Agriculture, Forestry and Fisheries as well as the Ministry of Posts and

Telecommunications were present. With strong endorsement from the Prime Minister, cooperation among different ministries involved was realized relatively easily in its early stage.

The government of Japan has been very sensitive toward reactions from abroad. As early as March 1986, MITI started its unofficial presentations abroad, and in early 1987 it conducted a survey among foreign scientists. Mr. Tamura, the Minister of International Trade and Industry presented an outline of plans for the HFSP in September, 1986, at the Japan-EC Symposium in Brussels. On April, 1987, a conference concerning the HFSP involving experts from the US and Europe was held in London.[31] It was entitled "London Wisemen's Conference for the Human Frontier Science Program" and was attended by scientists from Summit nations (Canada, France, Germany, Italy, Japan, U.K., and the USA)?[32]

Prime Minister Nakasone announced the HFSP at the Summit in June 1987. The HFSP was mentioned in the Economic Declaration at the Summit as follows:

> We welcome the initiative of the Human Frontier Science Program presented by Japan, which is aimed at promoting, through international cooperation, basic research on biological functions. We are grateful for the informal opportunities our scientists have had to take part in some of the discussions of the feasibility study undertaken by Japan. We would be pleased to be kept informed about its progress. (Economic Declaration No. 32.)

---

[31] This represents Japan's growing efforts to promote greater interactions with foreign governments in policy implementations.

[32] Participants from the US were Dr. Purnell Choppin (VP, Howard Hughes Medical Institute), and Dr. Howard Schneiderman (SVP, Monsanto Company).

Responses from other nations were at most lukewarm. After encountering a less than enthusiastic reception at the Summit in Italy, the HFSP is facing its first serious challenge within Japan. The HFSP needs to be clarified with a well-defined purpose. So far there have been neither concrete targets for the project nor a well-specified organizational mechanism.

C. Problems with HFSP

The Human Frontier Science Program is not yet well defined in terms of its goals and finances, as well as its relationship with the DNA sequencing effort.[33] The HFSP was initiated essentially for the purpose of international cooperation in the life sciences, and there is agreement among interested parties that the HFSP should include some mechanism to encourage such cooperation. However, there are neither concrete ideas for international cooperation, nor a framework for the transfer of technology to participating countries. Without a concrete plan to achieve truly mutually beneficial results among participating countries, the HFSP cannot gain necessary credibility.

This ambiguity of the goals and finance of the HFSP is causing further uncertainties. Given the very close ties in Japan between government and business, the goals and outcome of the Japanese project may be changed by pressure from the business community.

---

[33] Such a distinction is important for the U.S. in order to assess the final implementation of the Japanese sequencing efforts given interministrial differences, i.e. in commercialization of technologies.

Business leaders are, of course, pushing for a greater commercial return from government projects. The total spending for the HFSP is expected to be well above one trillion yen, and it is not likely to be financed entirely by the government itself (especially during the current budgetary constraints in Japan). If private companies in Japan are required to finance part of the project, the goals and outcomes may be affected since it is rational for private firms to expect a "return" on "investment."

Another uncertainty concerning the HFSP is that different government ministries, such as MITI and STA, have different attitudes toward the program. Even though the officially stated Japanese opinions are similar (at least as one observes in the US), different government agencies have quite different agendas. Though the HFSP was first discussed in Japan as a contribution to the "welfare of all human beings," the different attitudes of various government agencies toward the commercialization of technologies spun off from the project, may cause confusion and conflicts with other countries.

The HFSP faced internal disarray in Japan on the eve of the Summit in Italy. Though Prime Minister Nakasone has been one of the most earnest supporters of the HFSP, he reportedly expressed his reluctance to push for the HFSP at the Summit during his study meeting for the Summit. He maintained that the HFSP proposal drafted by his aides lacks both substance and concrete ideas. Countries can cooperate only for concrete targets such as the cure of AIDS and cancer, and without well-defined goals, he maintained,

it would be impossible to expect support from other nations.[34] His apparent change of opinion was affected by at least two factors. One is the not-so-supportive feedback from abroad, especially from the US, from a survey of foreign scientists.[35] The other factor is differences in opinions toward the HFSP among ministries, especially between those directly involved (MITI, STA) and those not directly involved (Ministry of Foreign Affairs (MFA), and Ministry of Finance (MOF)).[36] MFA is passive toward the HFSP, since it is concerned with foreign criticisms of the program. Similarly, MOF is not supportive with the Program--the Ministry is not eager to create an expensive project since Japan has been facing a serious governmental financial deficit. In fact, some MFA officials reportedly maintained that Prime Minister Nakasone's statement essentially meant that the HFSP should be either cancelled or totally modified.[37]

These ambiguities concerning the HFSP make it very difficult to analyze its relation to and its impacts upon the Japanese effort to sequence DNA. If the HFSP is approved formally, the HFSP will definitely affect the implementation of the sequencing effort since DNA sequencing is stated as one of the leading key technologies related to the HFSP. (See Figure C.) The following is a list of key technologies:

---

[34] Nihon Keizai, 6/5/87.

[35] Sankei Shinbun, 6/2/87.

[36] Ibid.

[37] Nihon Keizai, 6/6/87.

1) chemical determination of biological components.
2) Sequence analysis of DNA.
3) 3-dimensional structure determination of proteins.
4) measurement of microstructure of biological cells.
5) Non-invasive measurement of biological functions.
6) measurement of the dynamic structure of biological systems.
7) measurement of the function of biologican movement.
8) rDNA technology.
9) ultramicro manipulation.
10) biological cell culture
11) synthesis of biopolymers

I suspect that the DNA sequencing effort will be at least partially included in the HFSP. Though STA officials may insist that the sequencing effort should be independent of the HFSP and remain under STA control, the urgent need to make the HFSP more concrete may force the DNA sequencing effort to be a part of the HFSP in order to give it a relatively credible focus.

Today's Japanese sequencing effort is led by STA. Will MITI gain more control over the sequencing effort if the HFSP is implemented? I believe that MITI's involvement in the sequencing effort itself will be limited <u>even if</u> the HFSP is actually implemented. STA has taken a prime leadership role in the Japanese sequencing effort, and STA will definitely try to maintain its position. If in fact there are immediate commercial implications from the sequencing effort, MITI may be discouraged from becoming involved given the problems MITI is already facing due to the trade conflict. Though the exact relationship between the DNA sequencing effort and the HFSP is not yet clear, it is likely that the

sequencing effort will, at least for the time being, remain under the leadership of the STA.[38]

V. INTERNATIONAL COOPERATION

A. Scenarios of International Cooperation

In this section, I briefly present several scenarios or options for international cooperation, and the Japanese reactions to them. Some scenarios of international cooperation were discussed above in the section concerning repository/data base. One must not treat views presented below as the official Japanese positions regarding such proposals. Japanese attitudes are likely to change as they interpret feedback from abroad concerning their projects especially after their anouncement of the HFSP.

Scenario 1: Each country concentrates on a particular human chromosome or on particular organisms (other than human). Neither Japanese government officials nor scientists like this idea. They think the idea is not "interesting enough" since no one would like to sequence a "boring" chromosome. They maintain that the following scenarios (scenarios 2 and 3) are much more realistic means for cooperation. These are the separation of different parts of a genome to different countries (scenario 2), or the separation of different functions: sequencing, informatics, cloning, ordering clones, etc. (scenario 3).

---

[38] However the likelihood that the sequencing effort would be included in the HFSP increased after the Summit.

Scenario 4: The Japanese contribution would be to sequence the segments of DNA sent to Japan by Scientists in other countries. In another words, Japan would be a service station for the genome sequencing by establishing the Super DNA Sequence Center. This is similar to Professor Wada's Super Sequencing Center. He maintained:

> In conclusion, an international DNA analysis centre or centres equipped with super sequencing systems which are connected to a worldwide data-network should be developed, where most appropriate balance in man-machine cooperation throughout the whole DNA processing steps should carefully be examined by taking the input/output data-flow matching into consideration.[39]

Professor Wada (scientist) and STA officials disagree as to whether this is a viable option. Professor Wada believes that Japan is the most suitable place to establish such a Super Sequencing Center since the country has had much experience in mass production/quality control operation. However there are some criticisms to this plan. One of the issues is fairness: Japan may be criticized for focusing on the easiest parts and the last step, while leaving the difficult and expensive parts of the overall sequencing project to other countries. A second shortcoming may be that DNA sequence data can disclose details about vector design and cloning strategy. Scientists might be reluctant to send DNA to other laboratories for sequencing in the early stages of developing new cloning or mapping efforts because of fear that their new ideas

---

[39] A. Wada, "Many Small-Scale or a Few Large Scale DNA Sequencers?," 1987.

would be dispersed by having another party do the sequencing. An STA official criticized the scenario for a similar reason. If Japan's activity is limited to such a super sequencing center, he is concerned that there won't be enough contributions by the Japanese. The second criticism was also acknowledged as a potential problem by the STA official, however, he felt that it is important for Japan to avoid becoming just another "subcontractor" in science.

B. Cooperation in Science

Though we are facing many uncertainties today, I believe that it is safe to say that the Japanese are more than happy to consider any suggestions. Japan is deeply concerned with the present trade and technological conflicts with the U.S. Although Japan is suffering from "credibility problems," I believe the initial incentive to introduce the HFSP was to make sizable contribution to the world in basic understanding of the life sciences. Japanese scientists don't have credibility problems--the quality of Japanese science is well recognized elsewhere in the world.[40] Japanese scientists are eager to conduct research with American colleagues. However, it is the process of Japanese policy formulation, the role of bureaucrats, and the close ties between government and business in Japan, which foreign observers have sometimes questioned.

International cooperation is certainly not impossible. Japanese scientists have collaborated with scientists in the U.S. and in

---

[40] This point was raised by a reviewer.

Europe in various projects."[41] Many American scientists have established close friendship with Japanese scientists over years of close cooperation. However, one must realize although friendships within the scientific community may be sufficient to bring an informal "small" project to fruition, a "large" science project which requires commitments of more than two governments as well as scientists is a more complicated matter. In such a large international scale project, consensus among different governments concerning various issues, such as (1) the actual mechanism of technology transfer to participating countries to ensure equal benefit from the project, and (2) the role of government in actual commercialization, is required.

Realistically, however, it is not easy for the U.S. to propose such an international cooperation with Japan. Besides all the uncertainties and difficulties I have mentioned concerning Japan, there are many detailed arrangement the two countries would have to work out. For example, a clear cut principle must be established concerning policy in depositing clones, probes, or cell lines to be patented. Similarly what would be the protections on the disclosure or nondisclosure of DNA sequencing data?"[42] It is important for the U.S. and Japanese government to clarify these issues.

---

[41] One of the reviewers of the first draft raised this issue.

[42] This statement does <u>not</u> imply that Japanese are overly secretive. Instead I'm pointing out the current lack of guidelines concerning these legal issues (i.e. patentability of sequencing data). Patentability of DNA sequencing data has been discussed in the U.S., yet there is no consensus today.

VI. CONCLUSION

Though there are still many uncertainties concerning the Japanese sequencing efforts and its relationship with the HFSP, it should be stressed that it is essential for the U.S. to continue to carefully observe these developments in Japan. Uncertain interrelations between the government and business as well as inter-ministrial interactions will affect the actual implementation and outcome of the project. The possibilities are great for cooperation between the U.S. and Japan in DNA sequencing or in the HFSP, and for realizing a joint scientific breakthrough. Detailed discussions concerning issues of commercialization and mechanisms for technological transfers are called for, and actual cooperation should be conducted with great care. Without such detailed discussions, efforts to cooperate by well-meaning scientists from the two countries may be misguided and may end up in conflict.

REFERENCES

Agency of Industrial Science and Technology. AIST 1986 (Tokyo: Ministry of International Trade and Industry, 1986)

Agency of Industrial Science and Technology. Human Frontier Science Program (a memo distributed by AIST) (Tokyo: Ministry of International Trade and Industry, Jan. 1987)

BIDEC (Biotechnology Development Center). Japan Biotechnology Letters

BIDEC. Baiotekunorojii Kenkyushiensangyo Chosakenkyu Hokokusho (Report on Biotechnology R&D Equipment), (Tokyo: BIDEC, 1986)

Lewin, Roger. "Japanese Super-Sequencer Poised to Roll" in Science (3 April 1987)

Makino, Noboru ed. Haiteku Nippon'87 (Hightechnology Japan in 1987), (Tokyo: Sekaibunkasha, 1987)

Nikkei Sangyo Shinbunsha. Baio Bizinesu (Bio-business), (Tokyo: Nihon Keizai Shinbun, May 1987)

Nikkei Sangyo Shinbunsha. Baiotekunorojii Saizensen (Frontier of Biotechnology), (Tokyo: Nihon Keizai Shinbun, May 1984)

Prime Minister's Office. Organization of the Government of Japan 1986 (Tokyo: Institute of Administrative Management, 1987)

Saxonhouse, Gary. "Industrial Policy and Factor Markets: Biotechnology in Japan and the United States," in Hugh Patric ed., Japan's High Technology Industries (Seattle: University of Washington Press, 1986)

Science and Technology Agency. Kagaku Gijutsu Hakusho (White Paper of Science and Technology), (Tokyo: Ministry of Finance, Printing Office, 1987)

Science and Technology Agency. Deta Besu no Kodoriyou (High Level Uses of Data Base), (Tokyo: Ministry of Finance, Printing Office, 1985)

Science and Technology Agency. Kagaku Gijutsucho Nenpo (Annual Report of Science and Technology Agency), (Tokyo: Ministry of Finance, Printing Office, 1986)

Yoshikawa, Akihiro. The Japanese Challenge in Biotechnology: Industrial Policy (Berkeley, BIRE Working Paper, University of California, 1987).

Wada, Akiyoshi. "Japanese Super DNA Sequencer Project" in Science & Technology in Japan, (Tokyo: Three "I" Publications, April 15 1987)

Wada, Akiyoshi. "Many Small-Scale" or "A Few Large-Scale" DNA-Sequencers?" (a mimeo)

Wada, Akiyoshi. "Automated high-speed DNA sequencing" in Science (26 February 1987)

Wada, Akiyoshi. "DNA no Umie" (Toward Sea of DNA) in Kagaku (June 1987)

Current Development in Research in Relation to the Principal Biological Functions and Future Focal Points

The Human Frontier Science Program (April 1, 1987)

London Wise Men's Conference for the Human Frontier Science Program (April 1, 1987)

Appendix B.

LIST OF PARTICIPANTS

International Workshop

on

Automatic and High Speed DNA-Base Sequence

at

Hayashibara Biochemical Laboratory

July 7-9, 1987

West Germany

Dr. Lennart Philipson
Director EMBL Heidelberg
European Molecular Biology Laboratory
Heidelberg, Postfach 10.2209
F.R.G.

Prof. F. M. Pohl
Universitaet Konstanz
Postfach 5560
D-7750 Konstanz 1
F.R.G.

United Kingdom

Mr. J. K. Elder
Department of Biochemistry
University of Oxford
Oxford, OX1 3QU
U.K.

Dr. William M. Martin
The University of Manchester of Science and Technology
PO BOx 88, Manchester M60 1QD
U.K.

Prof. Edward M. Southern
Deparment of Biochemistry
University of Oxford
Oxford, OX1 3QU
U.K.

United States

Dr. George F. Cahill, Jr.
Vice President
Howard Hughes Medical Institute
Bethesda, Maryland

Dr. Charles Cantor
Department of Genetics and Development
College of Physicians and Surgeons of Columbia University
701 West 168th St.
New York, NY 10032

Dr. Anthony V. Carrano
Biochemical and Enviromental Science Division
Lawrence Livermore National Laboratory
Livermore, CA 94550

Dr. C. T. Caskey
Institute for Molecular Genetics
Baylor College of Medicine
Texas Medical Center
Houston, Texas 77030

Dr. George Church
Department of Genetics
Harvard Medical School
25 Shattuch Street
Boston, MA 02115

Prof. Walter Gilbert
Department of Biology
Harvard University
Cambridge, MA 02138

Dr. Walter Goad
Theoretical Biology and Biophisics
Los Alamos Naitonal Laboratory
Los Alamos, New Mexico 87545

Dr. Ed Hildebrand
Life Science Division
Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Prof. Leroy Hood
Division of Biology
California Institute of Technology
Pasadena, CA 91125

Dr. Michael Hunkapiller
Applied Biosystems
California

Keith McKenney
National Bureau of Standard
Chemical Thermodynamics (543)
Building 222, Room 3343
Gaitngrsburg, MD 20899

John Marenghi
EG & G Inc.
Watertown, MA

Dr. Richard Roberts
Cold Spring Harbor Laboratory
P.O. Box 100
Cold Spring Harbor, NY 11724

Dr. Lloyd M. Smith
Division of Biology 147-75
California Institute of Technology
Pasadena, CA 91125

Dr. Aki Yoshikawa
Berkeley Roundtable on the International Economy (BRIE)
215 Moses Hall
University of California, Berkeley
Berkeley, CA 94720

Japan

1. Business

Mr. Nobuyuki Baba
Toyo Soda Manufacturing Co., Ltd.

Mr. Yukihiro Eguchi
Mitsui Knowledge Industry Co.

Mr. Tsutomu Hamaoka
Fuji Photo Film

Mr. Hideki Kambara
Hitachi Ltd.

Mr. Shiro Kimura
Seiko Instruments Inc.

Mr. Yasutomi Mitomo
Toyo Soda Manufacturing Co., Ltd.

Mr.Keiichi Nagai
Hitachi Ltd.

Mr. Tetsuo Nishikawa
Hitachi Ltd.

Mr. Masaji Ogawa
Fuji Photo Film

Mr. Junichi Toei
Toyo Soda Manufacturing Co., Ltd.

Mr. Yoshinobu Uchida
Seiko Instruments Inc.

Academics

Mr. Masami Hasegawa
Institute of Statistical Mathematics

Dr. Katsumi Isono
Department of Biology
Kobe University

Dr. Hiroaki Kagawa
Faculty of Science
Okayama University

Prof. Minoru Kanehisa
The Institute of Chemical Research
Kyoto University

Dr. Chikao Kawabata
Computer Center
Okayama University

Dr. Yuji Kohara
Institute of Molecular Biology
Nagoya University

Dr. Kenji Ohyama
Faculty of Agriculture
Kyoto University

Ms. Fujiko Saka
The Institute of Phisical and Chemical Research

Dr. Yoshiyuki Sakai
Research Lalboratory for Genetic Information
Kyushu University

Prof. Nobuyoshi Shimizu
School of Medicine
Keio University

Dr. Eiichi Soeda
The Institute of Phisical and Chemical Research

Dr. Akira Suyama
The Technological University of Nagaoka

Prof. Hiromichi Tsurui
Faculty of Science
Tokyo Unoversity

Prof. Akiyoshi Wada
Faculty of Science
Tokyo University

FIGURE B

STRUCTURE OF THE HFSP

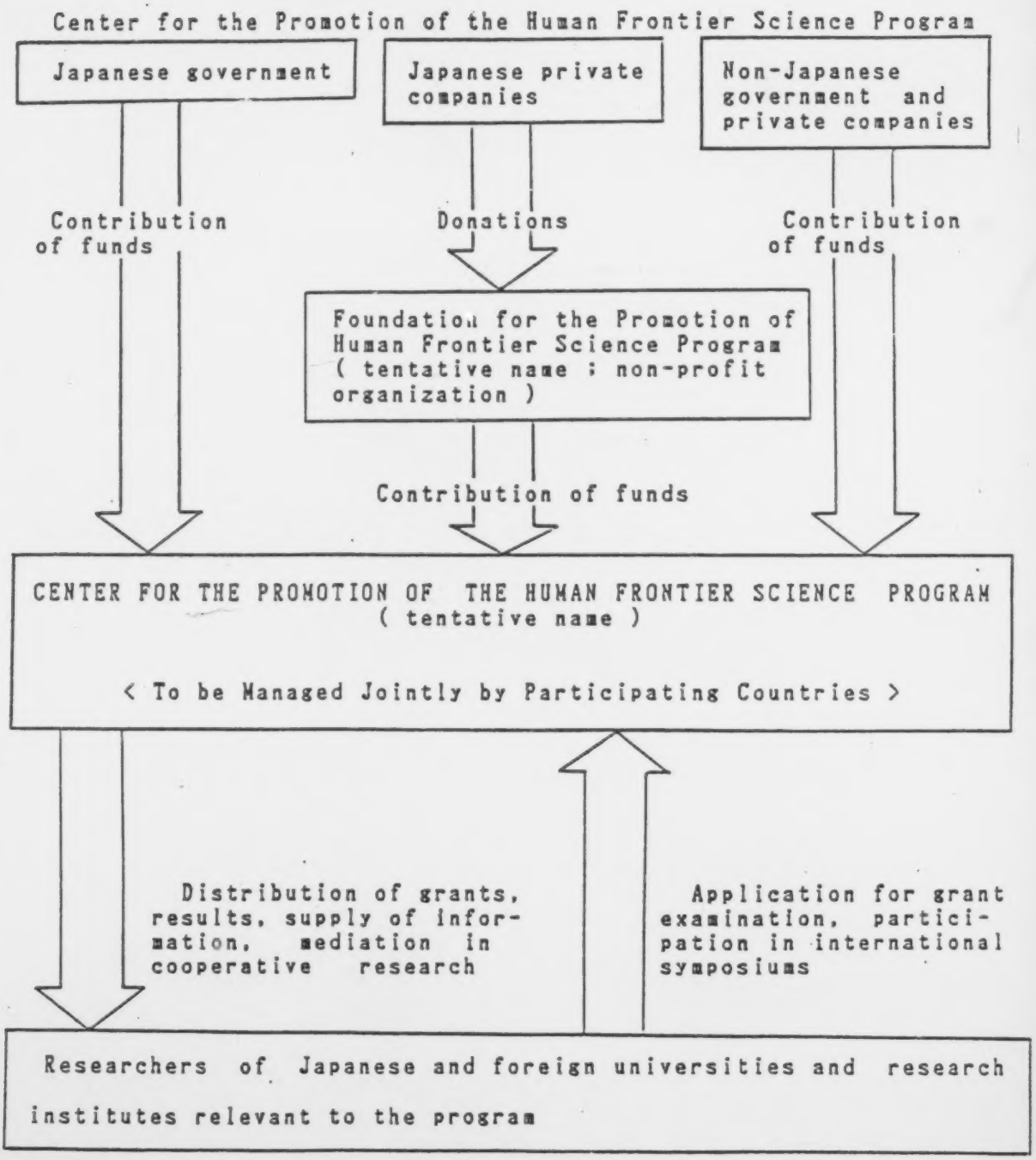

Center for the Promotion of the Human Frontier Science Program
Japanese government
Japanese private companies
Non-Japanese government and private companies
Contribution of funds
Donations
Contribution of funds
Foundation for the Promotion of Human Frontier Science Program (tentative name : non-profit organization)
Contribution of funds
CENTER FOR THE PROMOTION OF THE HUMAN FRONTIER SCIENCE PROGRAM (tentative name)
< To be Managed Jointly by Participating Countries >
Distribution of grants, results, supply of information, mediation in cooperative research
Application for grant examination, participation in international symposiums
Researchers of Japanese and foreign universities and research institutes relevant to the program

FIGURE C: RESEARCH SUBJECTS OF THE HUMAN FRONTIER SCIENCE PROGRAM

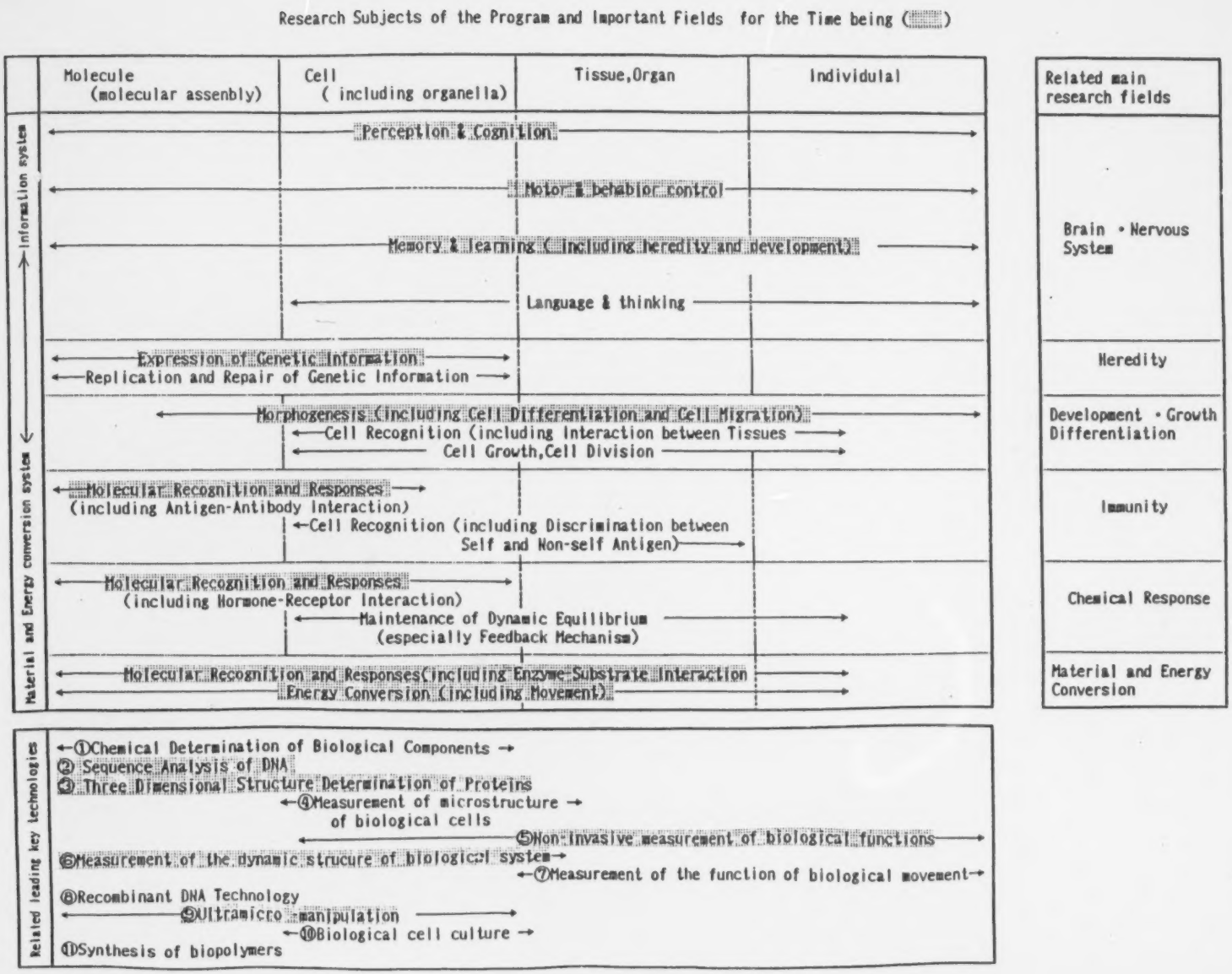

Research Subjects of the Program and Important Fields for the Time being (▒▒)
Molecule (molecular assenbly)
Cell (including organella)
Tissue,Organ
Individual
Related main research fields
Information system
Material and Energy conversion system
Perception & Cognition
Motor & behabior control
Memory & learning (including heredity and development)
Language & thinking
Brain • Nervous System
Expression of Genetic Information
Replication and Repair of Genetic Information
Heredity
Morphogenesis (including Cell Differentiation and Cell Migration)
Cell Recognition (including Interaction between Tissues
Cell Growth,Cell Division
Development • Growth Differentiation
Molecular Recognition and Responses (including Antigen-Antibody Interaction)
Cell Recognition (including Discrimination between Self and Non-self Antigen)
Immunity
Molecular Recognition and Responses (including Hormone-Receptor Interaction)
Maintenance of Dynamic Equilibrium (especially Feedback Mechanism)
Chemical Response
Molecular Recognition and Responses(including Enzyme-Substrate Interaction
Energy Conversion (including Movement)
Material and Energy Conversion
Related leading key technologies
①Chemical Determination of Biological Components
② Sequence Analysis of DNA
③ Three Dimensional Structure Determination of Proteins
④Measurement of microstructure of biological cells
⑤Non-invasive measurement of biological functions
⑥Measurement of the dynamic strucure of biological system
⑦Measurement of the function of biological movement
⑧Recombinant DNA Technology
⑨Ultramicro manipulation
⑩Biological cell culture
⑪Synthesis of biopolymers

TABLE A

Budget for Fiscal 1987

1. Feasibility study for fiscal 1987 and International Symposium for the Human Frontier Science Program (Autumn 1987, Tokyo)

---------- 150 million yen

2. Preparation of infrastructure relevant to the Human Frontier Science Program

(i) Study for the establishment of a biological engineering data base (data base on protein engineering and brain neutral circuits)

---------- 28.2 million yen

(ii) Study and research for the standardization of technical terms for the biological organism technology

---------- 18.8 million yen

Total: 197.0 million yen

APPENDIX A: MEMBERS OF THE HFSP FEASIBILITY STUDY COMMITTEE, ETC.

Members of the Feasibility Study Committee of the Human Frontier Science Program

	Name	Position	Remarks
1	Tetsuzo Akutsu	Associate Director, National Cardiovascul Center	artificial organs
2	Kei Arima	President, Japan Association of Industrial Fermentation. Emeritus professor, University of Tokyo	agriculture chemistry
3	Setsuro Ebashi	Managing Director, Institute for Molecular Science, Okazaki National Research Institutes	molecular physiology
4	Yoji Ikawa	Vice-Director, Tsukuba Lifescience Center, Institute of Physical and Chemical Research	molecular oncology
5	Morio Ikehara	Head, Protein Engineering Research Institute. Former head, Department of Pharmacology, Osaka University	protein engineering
6	Takemochi Ishii	Professor, Department of Engineering, University of Tokyo	Chairman of the Working Group for the Functions of the Conversion of Information by Organisms biological information processing
07	Masao Ito	Head, Department of medicine, University of Tokyo	physiology
8	Yoshikazu Ito	President, Toray Inc.	organic chemistry
9	Kunihei Kishi	Former Director-General, Agricultural Research Center	botanical physiology

	Name	Position	Remarks
16	Yoshinori Kato	Managing Director, NTT Basic Research Laboratories	electronics materials
11 ◎	Michio Okamoto	Chairman, Committee of Policy Matters, Counsil of Science and Technology. Former president, University of Kyoto	brain
12	Tairo Oshima	Professor, Laboratory of Chemistry for Natural Products, Faculty of Science, Tokyo Institute of Technology	bio-organic chemistry
13	Akio Sato	Managing Director, Fermentation Research Institute, Agency of Industrial Science and Technology, MITI	bioreactor
14	Tomio Tada	Professor, Department of Medicine, University of Tokyo	immunology
15	Saburo Tamura	Member, Science Council. President, Toyama Technical College	agricultural chemistry
16	Michiyuki Uenohara	Managing Director, NEC Corporation	electronics
17	Akiyoshi Wada	Professor, Science Faculty, University of Tokyo	Chairman of the Working Group for the Changing Functions of Biomaterials and Energy biophysics
18	Itaru Watanabe	Emeritus professor, Keio University	molecular biology

◎ ——— Chairman ○ ——— Vice-chariman

Members of the Working Group for the Feasibility Study
of the Human Frontier Science Program

1. The Working Group for the Changing Functions of Biomaterials and Energy

	Name	Position	Specialty
1	Goro Eguchi	Professor, National Institute for Basic Biology, Okazaki National Research Institutes	development biology
2	Akira Furuya	Director, Tokyo Research Laboratories, Kyowa Hakko Kogyo Co., Ltd.	microbiology, molecular biology
3	Tetsuya Hirano	Professor, Ocean Research Institute, University of Tokyo	marine biology and physiology
4	Yorinao Inoue	Chief Researcher, Institute of Physical and Chemical Research, Science and Technology Agency	photosynthesis
5	Yasuo Kagawa	Professor, Department of Biochemistry, Jichi Medical School	biochemistry, energy conversion by viomembrane
6	Masao Karube	Professor, Research Laboratory of Resources Utilization, Tokyo Institute of Technology	biosensor, development and utilization of biological resources
7	Yasuhisa Sakurai	Head, Institute of Biomedical Engineering, Tokyo Women's Medical College	artificial organs
8	Iwao Tabushi	Professor, Department of Synthetic chemistry, Faculty of Engineering, University of Kyoto	artificial cell
9	Toru Tokunaga	Director, Department of Cellular Immunology, The National Institute of Health	immunology

	Name	Position	Specialty
10	Keishiro Tsuda	Director of the First Department, Research Institute for Polymers and Textiles, Agency of Industrial Science and Technology, MITI	bioengineering, biopolymer
11 ◎	Akiyoshi Wada	Professor, Science Faculty, University of Tokyo	biophysics, genetic analysis
12	Jun Yamashita	Director, Department of Molecular Biology, National Institute of Agrobiological Resources	molecular biology, breeding
13	Kazuo Yoshida	General Manager, Research Center, Mitsubishi Chemical Industries Ltd.	catalytic process

◎ ------ Chairman

2. The Working Group for the Functions of the Conversion of Information by Organisms

	Name	Position	Specialty
1	Shozo Fujiwara	Senior Research Manager, Fundamental Research Laboratories, NEC corporation	magnetic memory
2	Humio Imamoto	Chief Researcher, Institute of Physical and Chemical Research, Science and Technology Agency	moleculor genetics
3 ◎	Takemoti Ishii	Professor, Department of Engineeneering, University of Tokyo	biological information processing

	Name	Position	Specialty
4	Yoshiaki Iwamura	Professor, Department of Medicine, Toho University	neurophysiology, somatosensory perception
5	Kazuhiko Kakehi	Head of Audio Visual Perception and Cognition, NTT Basic Research Institute	speech, auditory
6	Akio Maeda	Assistant Professor, Science Faculty, University of Kyoto	vision sensor, photoreception
7	Katsuhiko Mikoshiba	Professor, Institute for Protein Research, University of Osaka	biochemistry, neurotransmitor
8	Kuniaki Nagayama	Manager, Biometrology Laboratory, Japan Electron Optics Laboratory Co., Ltd	biometry
9	Takayuki Nakajima	Director, Information Sciences Division, Electrotechnical Laboratory, Agency of Industrial Science and Technology, MITI	information science (image processing and inference)
10	Kunihiko Obata	Professor, Department of Medicine, University of Gumma	pharmacology,
11	Hideo Sakata	Division-Director, Division of Physiology, Tokyo Metropolitan Institute for Neurosciences	neurophysiology, spatial perception
12	Tatsuro Shibuya	Professor, Institute of biological Science, University of Tsukuba	sensory physiology
13	Noboru Sugie	Professor, Department of Engineering, University of Nagoya	biological information processing, artificial intelligence

◎ ----- chairman